A MODEL BASED APPROACH TO NON-UNIFORM VOWEL NORMALIZATION
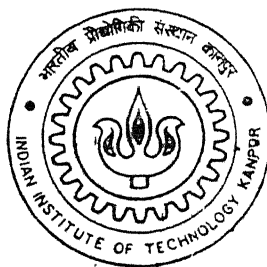
By

S. V. Bharath Kumar

DEPARTMENT OF ELECTRICAL ENGINEERING
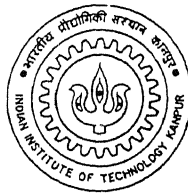
Indian Institute of Technology Kanpur

MARCH, 2002

A MODEL BASED APPROACH TO NON-UNIFORM VOWEL NORMALIZATION

*A Thesis Submitted
in Partial Fulfilment of the Requirements
for the Degree of*

Master of Technology

by

S. V. Bharath Kumar

to the

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

March, 2002

CERTIFICATE
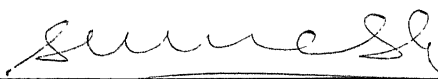
This is to certify that the work contained in the thesis entitled "*A Model Based Approach To Non-Uniform Vowel Normalization*", by *S. V. Bharath Kumar*, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

March, 2002

(Dr. S. Umesh)

Associate Professor,

Department of Electrical Engineering,

Indian Institute of Technology,

Kanpur.

Abstract

A model based vowel normalization procedure is proposed based on our study of the nature of relationships between formant frequencies of speakers. Conventionally, uniform scaling relationship between formant frequencies of speakers is assumed. In this thesis, we explore non-uniform scaling relationship between formant frequencies and then do appropriate speaker normalization for application in automatic speech recognition. The proposed model based vowel normalization procedure is independent of vowel class and is completely derived from Peterson & Barney and Hillenbrand *et al.* vowel formant databases. The frequency-warping necessary to do non-uniform vowel normalization using the model based procedure is similar to log-warp function. This method has been analysed using various cluster-discriminability measures, scatter plots and HMM-based vowel recognizers.

In this thesis, we also made a comprehensive study on the vowel normalization methods based on frequency dependent scaling of formant frequencies and scale-invariant transformation, each of which shows that the frequency-warping function required for normalization is a compromise between log-warp and mel-warp functions. Using separability measures like F-ratio and residual variance, the proposed method is found to be superior to Nordström & Lindblom's uniform scaling method and Fant's non-uniform normalization method. In addition, we have also compared the vowel-recognition performance of the proposed method with the other methods in a HMM-based recognizer. Using recognition accuracy as the performance measure, the proposed model based method is found to provide the best normalization for cross-gender cases.

Acknowledgements

At the outset, I deeply thank the Almighty for his sheer benevolence, for bestowing me with what I have.

I would like to express my profound gratitude to my thesis supervisor Dr. S. Umesh for his kind-hearted support, constant encouragement and valuable guidance throughout the course of my thesis work. I thank him for his concern towards my well-being and for showering on me his treasure of knowledge from both in and out of academic field. My deep regards to him for making my stay in the lab memorable by maintaining an informal atmosphere.

I would like to thank all my instructors at IIT, Kanpur, for their valuable guidance. I would like to thank my critic, mentor and lab mate Rohit for all his help which cannot be expressed in words. I specially thank my lab mate Rajesh for providing brain-storming discussions both in and out of academic arena throughout my stay at IIT, Kanpur. My association and experience with Rohit and Rajesh are unforgettable. I would like to thank Shafi for his valuable tips and warm friendship.

I thank *Pakki* Shankar, Manoj and Bhavani Shankar for providing me joyous moments in *baddy* court. I thank Andhra Sāmskritika Samiti (ACA) for providing me a "at home" experience at IIT, Kanpur. I thank all my DSP classmates (Anil, Harsha, Murali, Sampurna, Shashikant) and my beloved discussion pal, Rajesh for making my stay at IIT, Kanpur a memorable one. I would like to thank GE for providing me a scholarship covering all my expenses during the entire course period.

Finally, I wish to deeply thank my family members for being supportive in my every endeavour. I am grateful to Checha and Amma for their immeasurable love and support. Thanks to Checha, Amma and my sweet hearts Ranga and Hari without whom I wouldn't have been what I am.

S. V. Bharath Kumar

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Automatic speech recognition (ASR) system enables a computer (or a machine) to recognize words spoken by a person. Automatic recognition of speech by machine has been a goal of research for more than four decades. However, inspite of the glamour of designing an intelligent machine that can recognize the spoken word and comprehend its meaning, and inspite of the enormous research efforts spent in trying to create such a machine, we are far from achieving the desired goal of a machine that can understand spoken discourse on any subject by all speakers in all environments. The problem of ASR is dependent on many factors such as vocabulary size, speaker characteristics, accent, noise and channel characteristics. Hence, the whole problem can be tackled under two broad categories (1) Robustness to speaker variations and (2) Robustness to noise and channel effects. Assuming that the ASR system is robust towards noise and channel effects, the only major factor that affects its performance is the variability among speakers.

Depending on the speaker characteristics of the dataset used to train an ASR system, there are broadly two classes (1) Speaker Dependent (SD) and (2) Speaker Independent (SI) systems. Speaker dependent systems are trained from speech data collected from a single user, who is the sole user of the system. On the other hand, speaker independent systems are trained from speech collected from many different users. Typical applications of SD systems include desk-top applications, word processing, etc., while SI systems are typically used at public interfaces like airline interface system, telephone directory service, etc. where there are varied type of speakers. While the SI systems yield better recognition rates for speakers who are

not in the training dataset than speaker dependent systems, they are less accurate than adequately trained SD systems for a given speaker who has contributed to the training dataset. This degradation in the performance of SI systems over SD systems for a given speaker is mainly due to presence of large speaker variability in the training set of SI systems. Hence, it may be possible to achieve performances close to SD systems for a given speaker, if the variability in the training set is removed/reduced. The aim of speaker normalization techniques is to remove these speaker specific variabilities from the SI systems. These speaker variabilities can be due to physiological differences in the speech production apparatus or non-physiological factors like dialect, emotions, speaking idiosyncrasies etc.

A major source of the variability among speakers is attributed to the physiological differences in the vocal tract of the speakers. As an approximation, the vocal tract is assumed to be of uniform cross-section, in which case the speaker variability is directly related to the vocal tract length (VTL). It has been found that VTL variation causes scaling in the spectral domain [1] since the formant frequencies are inversely proportional to length of the tube [2]. Many normalization schemes, both linear scaling [1, 2, 3, 4] and non-linear scaling [5, 6] (of formant frequencies) have been proposed which compensate for this variability by re-scaling the frequency axis, resulting in substantial improvements in speech recognition performance [1, 2, 3, 4, 5, 6]. However, Fant [5] and others [6, 7] have shown that uniform/linear scaling of formant frequencies is a very crude approximation and that the formant scaling is non-linear and is phoneme dependent.

In this thesis, we have attempted to model these non-linearities in scaling as a function of frequency alone and have decoupled it from phoneme dependence unlike other methods. We have made a study of relationships between formant frequencies of speakers to understand the nature of non-linearity present between them. Based on this study, we have developed a model for the non-linear relation between the formant frequencies and applied it for vowel normalization. The frequency-warping function calculated based on the model we developed is found to be close to log-warp function.

In addition, we have also made a comprehensive study related to non-linear scaling of formants in which we estimate an improved frequency-warping function as compared to [8] which is a compromise between log-warp and mel-warp functions. We have also obtained a similar warping function based on modification of

scale-invariant transformation [7]. We have used different analysis methods such as formant data analysis, scatter plots and HMM-based vowel recognizers to compare the performance of the proposed model based vowel normalization procedure to that of other similar techniques.

The thesis is organized as follows. In Chapter 2, the motivation for non-linear scaling of formants is provided while discussing Nordström & Lindblom's [1] linear scaling method, Fant's [5] non-uniform normalization and frequency dependent scaling [6, 8] method. In Chapter 3, the proposed model based vowel normalization is presented. In Chapter 4, we compare the performance of our proposed method with the methods discussed in Chapter 2 in terms of residual variance, F-ratio and scatter plots. In Chapter 5, we present our comprehensive study in modelling the non-linear scale factor to obtain a frequency-warping function. In Chapter 6, the performance of these normalization schemes are evaluated, analysed and compared with respect to that of other similar techniques, using percentage accuracy as performance measure, by incorporating them into a HMM-based vowel recognizer. Finally, in Chapter 7, using all the experimental results, conclusions were drawn about the effectiveness of the proposed model based normalization approach in a recognizer/classifier framework.

Chapter 2

Vowel Normalization by Frequency Dependent Scaling

One of the major factors affecting speech recognition is the speaker dependence of the speech signal. It is a well known fact that because of the differences in vocal tract dimensions, two speakers may produce vowels that sound similar although they have very different formant values, and they may also produce vowels that sound different but which have remarkably similar formant values [9]. Generally, as a first-order approximation, the vocal tract shape is assumed to be a tube of uniform cross-section. Hence, difference in lengths lead to difference in formant values. An average adult male has a vocal tract length (VTL) of around 17cm, while the average female VTL measures around 14.5cm. The first-order effect of the difference in VTL is the scaling of the frequency axis, i.e., on an average the formants of an average female speaker are scaled up by 20% with respect to that of an average male speaker, with the differences most severe in vocal tract configurations in open vowels. Hence, it is commonly assumed that differences in formant patterns between male and female speakers are related by a pure scale factor which is inversely proportional to VTL [2, 10]. Different normalization procedures have been proposed in literature [1, 2, 5, 11] which counteracts the effect of varied vocal tract lengths.

2.1 Nordström & Lindblom Method : Simple Linear Scaling

Nordström & Lindblom [1] have proposed a simple normalization procedure based on an estimate of the speakers' average VTL in *open* vowels as determined from the measurement of the third formant F_3. As a support of their procedure, they demonstrated a substantial reduction of the male-female-child differences in the Peterson & Barney [9] database on American English vowels. In their procedure of uniform/linear scaling, the formant frequencies of the subject to be normalized are simply to be divided by the factor

$$\alpha = \left(1 + \frac{k}{100}\right) = \frac{F_{3_{sub}}}{F_{3_{ref}}} = \frac{l_{ref} + 1}{l_{sub} + 1} \qquad (2.1)$$

where k is the scale factor in percentage, $F_{3_{sub}}$ and $F_{3_{ref}}$ are the *average* F_3 of *open* vowels (vowels with F_1 greater than 600Hz) of the subject and the reference *"male"* speaker, l_{sub} and l_{ref} are the VTL's associated with the subject and the reference speakers respectively.

As mentioned before, the uniform tube is only a first-order approximation to the vocal tract shape, resulting in a uniform scaling to do the normalization. But in general, the formant frequency locations (in $F_1 - F_2 - F_3$ plane) for vowels are affected by three factors: the *effective* length of the pharyngeal-oral-tract, the location of constrictions along the tract, and the narrowness of the constrictions [12, 13]. Simple linear scaling neglects both the location of the constrictions and the vocal tract shape. Figure 2.1 shows that the simple linear scaling is a function of both formant number and vowel category.

2.2 Non-Linear Scaling

Fant [5] has suggested a non-uniform method of simple scaling procedure by modifying the correction factor, k as a function of both formant number and vowel category. With this non-uniform normalization, Fant showed a substantial reduction in speaker differences between male and female than the simple linear scaling as proposed by Nordström & Lindblom.
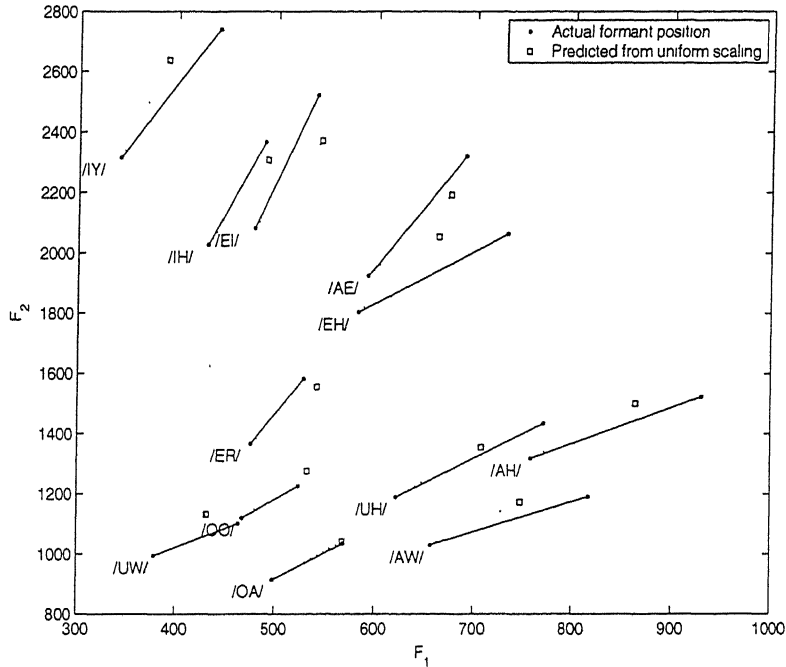
Figure 2.1: Deviations from linear scaling.
Figure shows the actual values of (F_1, F_2) points for average male and average female speakers for various vowel categories in the Hillenbrand et al. [14] database on American English vowels. Dashed line indicates predicted (F_1, F_2) point, with a linear scaling $\alpha = 1.14$. Variation in the distances between predicted and actual points, over vowel categories is evident. F_1 & F_2 are in Hz.

Fant calculated the reference scale factor, $k_{n\mathcal{F}}^j$ between the average female and the average male (i.e. the reference speaker) for the nth formant of the jth vowel class. Interested reader is referred to [5, 8] for the study of the formant specific weighting factors for different genders as defined by Fant. In non-uniform normalization procedure, Fant calculated the factor k for the average female to be 17, with average male being the reference speaker using a method slightly different from Eq. (2.1) and by using 6 to 8 different vowel databases of different languages. Apart from using the k (let us redefine it as k_{open}) as defined by Nordström & Lindblom, Fant also used the k value determined from $(F_2 F_3)^{\frac{1}{2}}$ of the front vowel /IY/, with 0.5 weighting, using the formula defined in Eq. (2.1). Thus the scale

factor that Fant used for non-uniform normalization procedure is given by,

$$k = \frac{2k_{open} + \frac{1}{2}(k_{2\,/IY/} + k_{3\,/IY/})}{3} \qquad (2.2)$$

Fant's non-uniform normalization for any particular adult subject speaker is given by the weighting of $k^j_{n\mathcal{F}}$ with the ratio of subject's particular k to the k = 17% of the average female speaker with respect to average male reference speaker i.e.

$$k^j_n = k^j_{n\mathcal{F}}\left(\frac{k}{17}\right) \qquad (2.3)$$

For the child speaker, Fant proposed the following non-uniform normalization scheme.

$$k^j_n = k^j_{n\mathcal{F}}\left(\frac{24}{17}\right) + (k - 24) \text{ for } k > 24 \qquad (2.4)$$

Eq. (2.3) and Eq. (2.4) represent the best prediction of the subject's scale factor for a particular formant of a particular vowel.

2.2.1   Experiments and Results

Earlier experiments [8] on Fant's non-uniform normalization method have showed a better vowel normalization for the average female as reference speaker than the average male as reference. So, in our experiments related to Fant's approach, the average female was chosen as the reference speaker. This selection provides a common normalization formula for both adult and child speakers given by

$$k^j_n = k^j_{n\mathcal{M}}\left(\frac{k}{\varphi}\right) \qquad (2.5)$$

where φ is the scale factor of an average male with respect to average female reference speaker calculated using Eq. (2.2). Based on Fant's approach, φ was calculated to be -14.65 for Peterson & Barney (PnB) database and -12.18 for Hillenbrand (HiL) database respectively. In the calculation of φ, (/AA/, /AE/, /EH/) and (/AE/, /AW/, /EH/) were considered as open vowels for PnB and HiL databases respectively. The subscript $\mathcal{M}$ in the notation $k^j_{n\mathcal{M}}$ is used to emphasize that the scaling is for the average male subject with respect to average female speaker as reference. Table 2.1 shows the $k_{n\mathcal{M}}$ values calculated using Fant's method for PnB and HiL databases.

2.3 Frequency Dependent Scaling Method

The main motivation behind this work was Fant's non-uniform normalization procedure. Basically, this work was motivated by the idea to apply Fant's non-uniform normalization method to reduce the inter-speaker difference thus aiding in speaker-independent speech recognition. Fant's non-uniform normalization scheme, though is definitely better than the simple uniform scaling, it cannot be directly applied for speaker-normalization since it requires knowledge of the vowel category and the formant number. The basic idea behind Frequency Dependent Scaling (FDS) method [5, 6] is to model the weighting factor $k_{n\mathcal{M}}$, as a function of frequency alone, thus making it context independent (i.e. independent of vowel category) and formant independent. This algorithm should do away with the need for *apriori* knowledge about the vowel category, while at the same time should do better than simple linear scaling.

2.3.1 Frequency Dependent Scaling Factor, Γ_f

The weighting factor, $k_{n\mathcal{M}}$ is a function of both formant number and vowel category. The $k_{n\mathcal{M}}$ value shown in Table 2.1, calculated using Fant's method was averaged over vowel category and formant number, for the respective databases to obtain a frequency dependent scaling factor, γ_f, which is purely a function of frequency [6]. The modelling of the weighting factor $k_{n\mathcal{M}}$ as a function of frequency alone was essentially done by plotting $k_{n\mathcal{M}}$ for each formant number and vowel as a function of subject's formant frequency. This was done for all speakers in the database and the averaging was done along the frequency axis over small bands of 100Hz width. γ_f represents the frequency dependent scale factor in a given 100Hz band. A vector of frequency specific scale factors, γ_f which are independent of formant number and vowel category was obtained. We denote this frequency dependent scale factor array as Γ_f. The subscript f shows that the parameter is frequency dependent. A plot of Γ_f is shown in Figure 2.2 for PnB and HiL databases, where each stem corresponds to the value of γ_f over a 100Hz band. The normalization scheme is given by

$$k_f = \gamma_f \left(\frac{k}{\varphi} \right) \tag{2.6}$$

Formant scale factors	PnB			HiL		
$k_{n\mathcal{M}}(\%)$	$-k_{1\mathcal{M}}$	$-k_{2\mathcal{M}}$	$-k_{3\mathcal{M}}$	$-k_{1\mathcal{M}}$	$-k_{2\mathcal{M}}$	$-k_{3\mathcal{M}}$
/AA/	17	11	12	-	-	-
/AE/	23	16	15	14	17	13
/AH/	17	15	14	18	14	11
/AO/	03	09	12	-	-	-
/AW/	-	-	-	20	14	10
/EH/	13	21	17	21	13	12
/EI/	-	-	-	11	17	12
/ER/	03	17	14	10	14	11
/IH/	11	19	16	12	14	12
/IY/	14	18	11	22	15	11
/OA/	-	-	-	12	12	13
/OO/	-	-	-	11	09	13
/UH/	07	12	16	19	17	12
/UW/	19	09	16	18	10	13

Table 2.1: Formant and vowel specific scale factors, $k_{n\mathcal{M}}$

Here '-' denotes that the corresponding vowels are not present in the respective databases. PnB refers to Peterson & Barney database and HiL refers to Hillenbrand database.

The above normalization procedure, in its present form, is applicable only to discrete formant patterns, as Γ_f is actually an array of weighting factors. But the state of the art modern day speech recognizers make use of continuous spectral patterns. Hence for FDS method to be implemented on a recognizer/classifier, the normalization procedure defined in Eq. (2.6) need to be extended for continuous spectral patterns. To modify Γ_f to be a continuous function of frequency, a simple curve was fitted using TableCurve2D to the array of weighting factors against frequency shown as a stem plot in Figure 2.2. Since the continuous function which was obtained is not exact, we term it as an approximate scaling function, $\gamma(f)$. The approximate scaling function along with the stem plot is shown in Figure 2.3 for both PnB and HiL databases. Table 2.2 shows the equations of the curvefits for frequency dependent scale factors for PnB and HiL databases. Since the scale factor defined
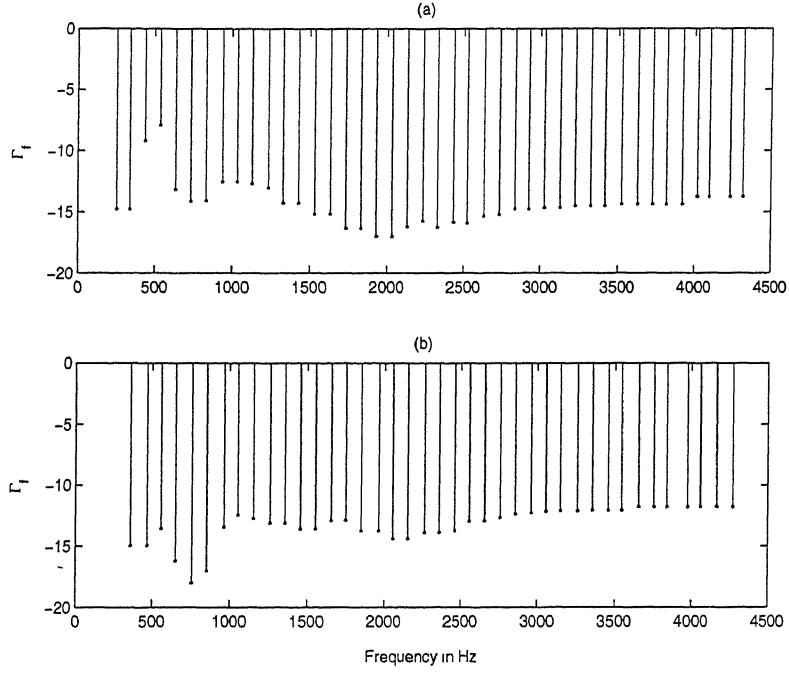
Figure 2.2: Frequency dependent scale factors, Γ_f

Figure shows the scaling factors in percentage for male group with female as the reference speaker for (a) PnB database and (b) HiL database

in Eq. (2.6) is frequency dependent, replacing k defined in Eq. (2.1) by k_f, the frequency dependent scale factor is obtained as

$$
\begin{aligned}
\alpha(f) &= 1 + \frac{k_f}{100} = 1 + \frac{\gamma_f}{100}\frac{k}{\varphi} \\
&\simeq 1 + \frac{\gamma(f)(\alpha - 1)}{\varphi}
\end{aligned}
\tag{2.7}
$$

With this approach, though there is no context dependence, yet we need to explicitly calculate the scale factor for each speaker to do normalization. Eq. (2.7) shows that $\alpha(f)$ is not only frequency dependent but also speaker dependent because of the presence of factor α or k. To achieve speaker normalization, one way is to explicitly estimate the scale factor, α [3, 4, 5] for each speaker and use it to normalize him with respect to some reference speaker. The other way is to use the knowledge of scaling function, $\gamma(f)$ so that a *universal* frequency-warping function leading to scale invariance [7, 10] can be developed. This will be of immense importance

Database	$\gamma(f)$
PnB	$-15 + 4.63e^{-f/723.21}$
HiL	$-16.53 + 0.04f^{0.59}$

Table 2.2: Equations of curvefits for $\mathbf{\Gamma}_f$.

$\gamma(f)$ *is the closed form equation for* $\mathbf{\Gamma}_f$. *The curvefits were obtained by using Table-Curve2D package.*

in deriving speaker independent robust features. This motivated us to derive a frequency warping function for frequency dependent scaling method.

2.4 Frequency-Warping Function Based on Frequency Dependent Scaling Method

Consider two speakers $\mathcal{A}$ and $\mathcal{B}$, whose spectras are related by

$$S_{\mathcal{A}}(f) = S_{\mathcal{B}}(\text{g}(\alpha_{\mathcal{A}\mathcal{B}}, f)) \tag{2.8}$$

where $\text{g}(\alpha_{\mathcal{A}\mathcal{B}}, f)$ is some function that involves speaker dependencies through the first argument. Let $f' = \text{g}(\alpha_{\mathcal{A}\mathcal{B}}, f)$. Our aim was to determine g(.) so that $\text{H}(S_{\mathcal{A}}(f)) = \text{H}(S_{\mathcal{B}}(f))$, where $\text{H}(.)$ denotes some mapping from f-domain to some domain, say η. The non-linearity $\text{g}(\alpha_{\mathcal{A}\mathcal{B}}, f)$ has been modelled in many parametric forms [7, 15, 16]. We modelled it as

$$f' = \text{g}(\alpha, f) = \alpha(f)f = \alpha^{\beta(f)}f \tag{2.9}$$

where α is the subject's scale factor with respect to a reference speaker, which is frequency independent and $\beta(f)$ is only frequency dependent and is independent of speaker. $\beta(f)$ captures the non-linearity in scale factor. Eq. (2.9) can be modified as

$$\log(f') = \beta(f)\log(\alpha) + \log(f) \tag{2.10}$$

$$\frac{\log(f')}{\beta(f)} = \log(\alpha) + \frac{\log(f)}{\beta(f)} \tag{2.11}$$

Define

$$\nu = \text{W}(f) = \frac{\log(f)}{\beta(f)} \tag{2.12}$$
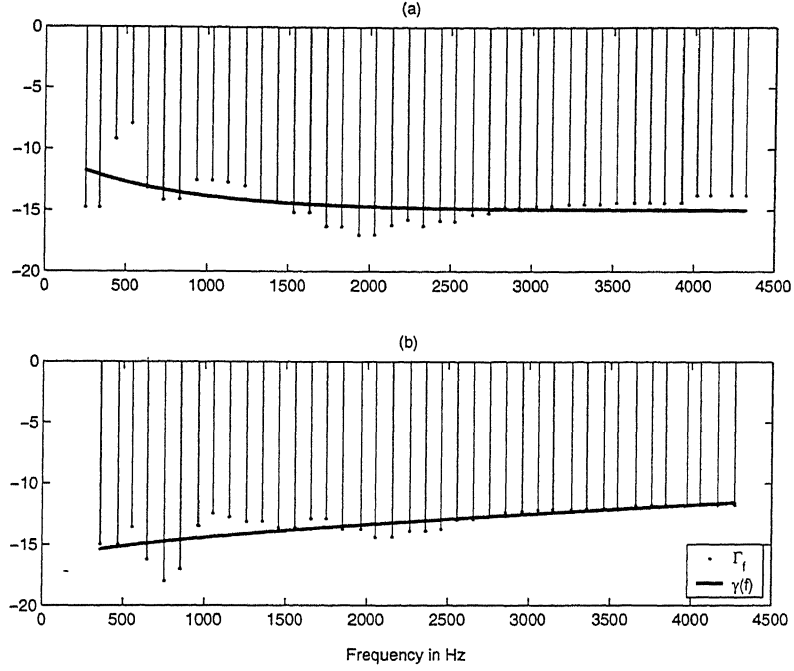
Figure 2.3: Frequency dependent scaling factor, Γ_f and frequency dependent scaling function, $\gamma(f)$

Figure shows the frequency dependent scaling function, $\gamma(f)$ obtained by curvefitting Γ_f for (a) PnB database and (b) HiL database.

Assuming $\beta(f') \simeq \beta(f)$, we get

$$\nu' = \nu + \log(\alpha) = \nu + \text{constant shift} \qquad (2.13)$$

where ν is the warped domain and $W(f)$ is the frequency warping function. Eq. (2.13) shows that the spectras in the warped domain are translated versions of one another. The magnitude of the Fourier transform of these warped spectral patterns are invariant to translations, leading to scale invariant features, of real speech signals. For the given model, frequency-warping function can be derived as,

$$\alpha(f) = 1 + \frac{\gamma(f)(\alpha - 1)}{\varphi} = \alpha^{\beta(f)} \qquad (2.14)$$

hence,

$$\beta(f) = \frac{\log(1 + \frac{\gamma(f)(\alpha-1)}{\varphi})}{\log(\alpha)} \qquad (2.15)$$

Eq. (2.15) is valid for all values of α, as $\beta(f)$ is assumed to be speaker independent.
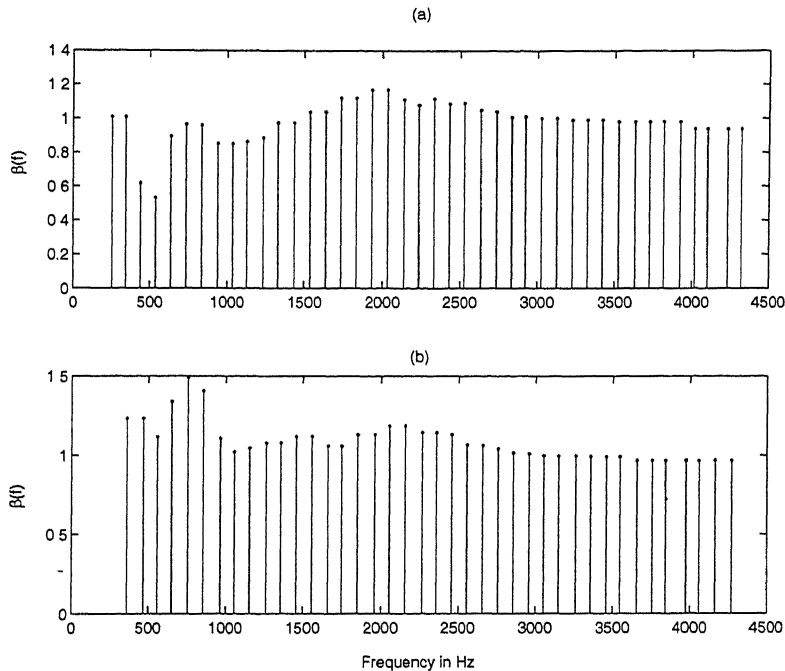
Figure 2.4: $\beta(f)$ derived from frequency dependent scaling factors, Γ_f
$\beta(f)$ was calculated using Γ_f, instead of $\gamma(f)$ to avoid the curve-fitting errors. Figure
shows the plot of average $\beta(f)$ for (a) PnB database and (b) HiL database.

2.4.1 Experiments and Results

In our experiments with FDS method, we obtained $\gamma(f)$ by fitting a simple curve to
Γ_f values. Figure 2.3 shows that the curvefit, $\gamma(f)$ is not accurate and infact, it is
only 60% accurate. Hence in order not to introduce curve-fitting errors, we carried
out our experiments using Γ_f itself instead of $\gamma(f)$. Though our assumption is that
$\beta(f)$ in only frequency dependent, actually, $\beta(f)$ will be different for different values
of α which is obvious from Eq. (2.14). Values of $\beta(f)$ obtained for all speakers in
the database were averaged over and an *average* $\beta(f)$ was calculated for both PnB
and HiL databases. Since $\beta(f)$ was calculated from Γ_f, it is actually not a function
of frequency but an array of factors dependent on frequency. Figure 2.4 shows the
average $\beta(f)$ obtained for PnB and HiL databases.

The frequency-warping function, $W(f)$ defined in Eq. (2.12) is not correct
as the assumption $\beta(f') \simeq \beta(f)$ is not valid. Since finding the closed form solution

Database	W(f)
PnB	$-3400.68 + 649.40 \log(f)$
HiL	$-1350.19 + 50.03 \left(\log(f)\right)^2$

Table 2.3: Equations of curvefits for discrete warping function, $W_i(f)$.
$W(f)$ *is the closed form equation for the discrete warping function,* $W_i(f)$. *The curve-fits were obtained by using TableCurve2D package*

for frequency-warping function is difficult, we discretized the warping function over a set of frequency bands and calculated the warping function for each band. A similar kind of situation arises in Section 5.3.1, where a discrete warping function was calculated, the warping parameters being calculated in a different way. Given the warping parameters, the methods used to compute discrete warping function were exactly the same. We present the method of finding the discrete warping function with full details in Section 5.3.1. Here, we present the details about the warping function that was calculated using the method described in Section 5.3.1.

In short, the method is as follows. Let us divide the frequency axis into N logarithmically equi-spaced regions. $\beta(f)$ was discretized into β_i's, where β_i is the value of β in i^{th} frequency region. β_i's were calculated by averaging the values of $\beta(f)$ that lie in i^{th} frequency region. The value of N has to be so chosen that $\beta(f)$ and β_i's should not be very different. We chose $N = 10$. Figure 2.5 shows the plot of $\beta(f)$ and β_i's. The warping function for each band was calculated as

$$W_i(f) = \frac{\log(f)}{\beta_i} \, , \, 1 \leq i \leq N \tag{2.16}$$

The closed form solution to the frequency-warping function, $W(f)$ was obtained by fitting a curve to the discrete warping functions, $W_i(f)$, $i = 1, 2, \cdots, N$. Table 2.3 shows the equations of curvefits for the discrete warping function, $W_i(f)$ for PnB and HiL databases. Figure 2.6 shows the actual warping function, $W_i(f)$ and its curvefit, $W(f)$ for PnB and HiL databases. This is plotted to show how close the curvefit is to $W_i(f)$ as $W(f)$ is the one that is actually required while implementing the normalization method on continuous spectral patterns. Figure 2.7 shows the warping functions derived using FDS method along with log-warp and mel-warp
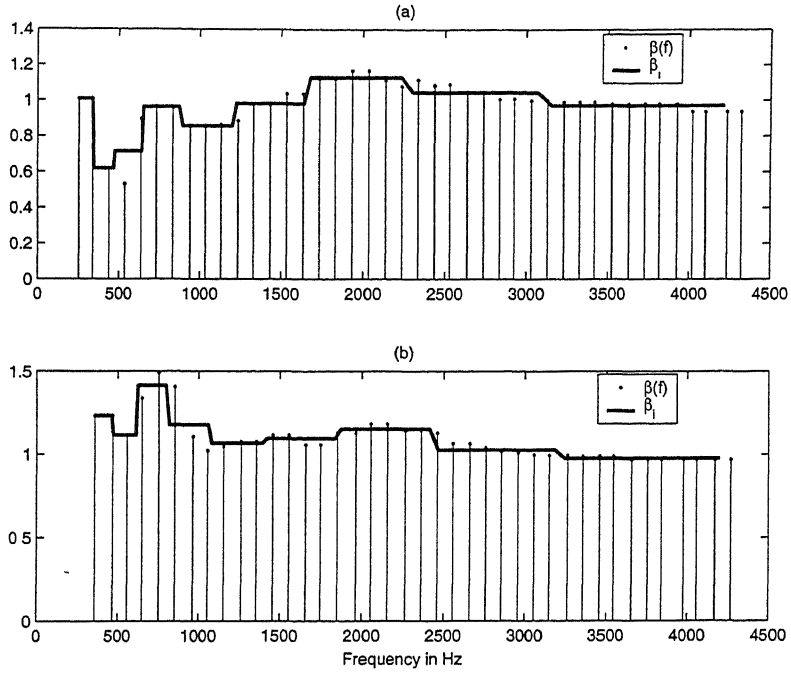
Figure 2.5: $\beta(f)$ and its discrete version, β_i

Figure shows average $\beta(f)$ along with its discrete version, β_i for (a) PnB database and (b) HiL database. It is obvious to see that β_i provides a good representation of $\beta(f)$ which helps in deriving the discrete warping function, $W_i(f)$, Eq. (2.16).

functions, the latter [17] being defined by

$$W_{mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (2.17)$$

It is very interesting to note that the warping function derived using FDS method lies between log-warp and mel-warp curves. $W(f)$ follows log-warp curve at low frequencies (< 500Hz) and mel-warp curve at high frequencies (> 3500Hz). The warping function derived out of HiL database is more closer to mel-curve at high frequencies than the one derived from PnB database. Log-warping refers to simple linear or uniform scaling of formant frequencies of the speakers. Mel-warp function was applied in speech recognition not from the speaker normalization point of view but from psychoacoustic view point. Since human ear behaves on mel-scale [18], mel-warping function is used in speech recognition to emulate the human ear. Now, our experiments with FDS method has revealed a frequency-warping function derived
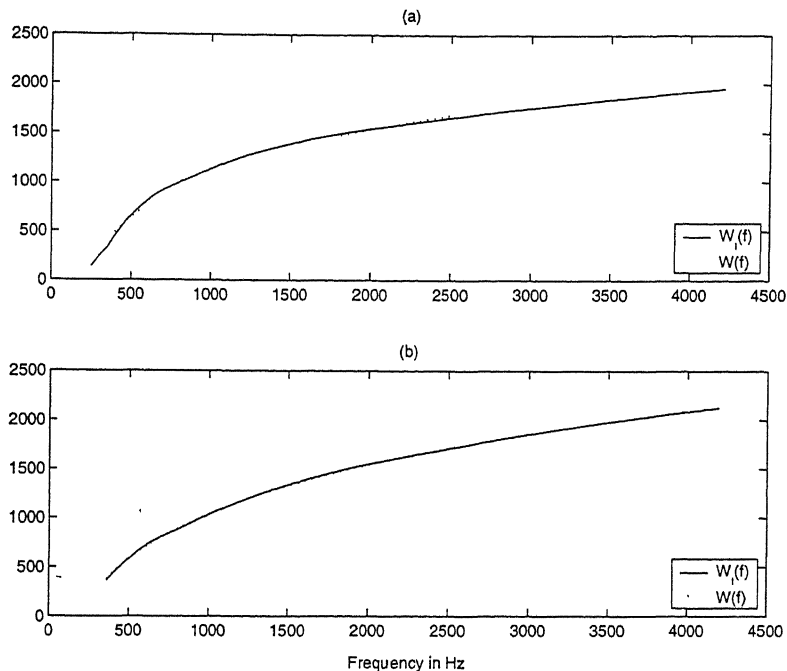
Figure 2.6: Warping function, $W_i(f)$ and its closed form approximate, $W(f)$

Figure shows the warping function, $W_i(f)$ and its closed form approximate, $W(f)$ given in Table 2.3 *for (a) PnB database and (b) HiL database. The curvefits to $W_i(f)$ were the best fits obtained from TableCurve2D.*

using speech data alone, which behaves mel-like at higher frequencies and log-like at lower frequencies and acts as a compromise between the two in the middle region, which is the region of interest in speech recognition.

2.5 Summary

The differences in vocal-tract dimensions among the speakers is one of the major factors affecting speech recognition. The first-order approximation of the vocal tract dimension to a uniform tube results in uniform scaling of formant frequencies. Nordström & Lindblom's and Fant's methods of normalization were presented. A frequency dependent scaling method was proposed, which is formant and context independent unlike Fant's non-uniform normalization method. The non-linearity of the scaling function obtained from frequency dependent scaling method was mod-
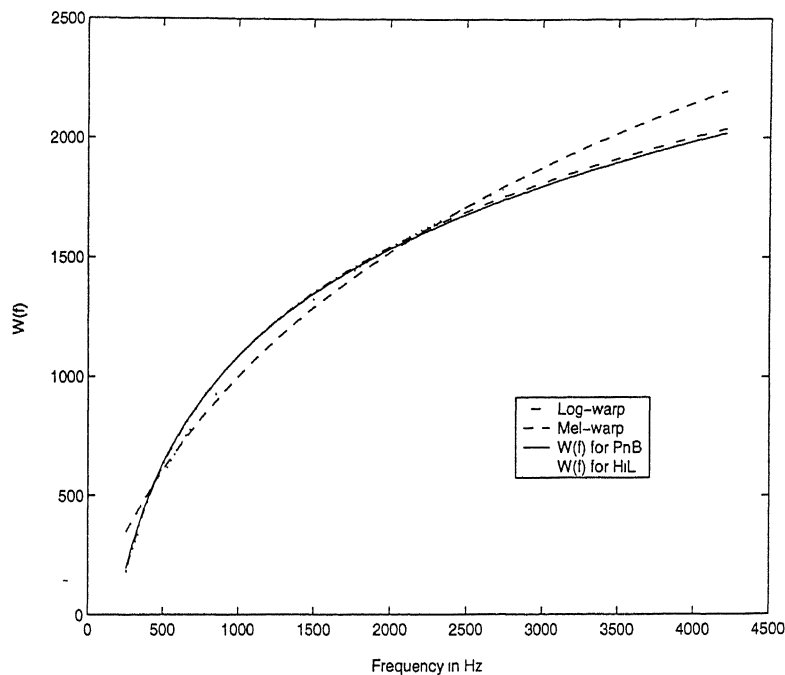
Figure 2.7: Comparison of warping function derived from frequency dependent scaling method with log-warp and mel-warp functions.

Figure shows the warping functions derived using FDS method along with log-warp and mel-warp functions. Log-warp function refers to uniform scaling of the formant frequencies of speakers. Mel-warp function is derived from psychoacoustic studies. It is interesting to note the similarity of these warping functions, though they are derived from entirely different studies.

elled and a warping function was derived out of the speech data which resembles log-warp function at low frequencies ($< 500Hz$) and mel-warp function at high frequencies ($> 3500Hz$). An interesting point was made in deriving a frequency-warping function that stands as a compromise between mel-warp and log-warp functions.

Chapter 3

Study of Relationships Between Formant Frequencies of Speakers

3.1 Motivation

For last five decades, a lot of research has been carried out to solve the problem of speech recognition and a large amount of understanding has been developed in this process, but the complete solution of the problem still remains elusive. The problem of speech recognition is very broad and we restricted ourselves to speaker independent speech recognition. Inter-speaker variation is a major factor that affects the performance of speaker independent speech recognition. Human ear is *the* system that has very high speech recognition accuracy than any modern day recognizer. A lot of research has undergone in the field of psychoacoustics leading to the understanding of human auditory mechanism [18, 19]. One can ask a very basic question. "Why don't we embed the psychoacoustic properties of ear into the machine". Thus, the knowledge gained in the field of psychoacoustics was employed into the recognizers. This lead to the cropping of terms "Mel Scale" [18] and "Bark Scale" [20] into the parlance of speech recognition. Though the recognition rates were improved, still there remains vacuum to be filled up.

Human beings are able to handle the variability of speech very well, which is not the case with the machines yet. Here one interesting point to note is that the variability in the speech data (due to the physiological differences in the speech production system) is sort of nullified by the human auditory mechanism. The problem

18

of making the machines recognize better was attacked initially from psychoacoustic view point and not from the point of view speaker normalization. In order to understand and parameterize the variability in speech, one might be motivated to ask, "Given any two speakers, what is the relationship between them in terms of speech related parameters". The solution to this question will provide an insight into the speech production mechanism itself. Given a set of speakers, if we can find a *universal relation* relating all of them to a reference speaker, it will be very helpful in normalizing all the speakers to a single reference speaker thus aiding in speaker independent speech recognition.

3.2   Model Based Normalization

We propose the following model relating the formant frequencies of the subject speaker and the reference speaker as

$$F_{\mathcal{R}} = a_{\mathcal{R}\mathcal{S}} \left(1 + \frac{F_{\mathcal{S}}}{b} \right)^{c} \qquad (3.1)$$

where $F_{\mathcal{R}}$, $F_{\mathcal{S}}$ are the formant frequencies of the reference speaker, $\mathcal{R}$ and the subject speaker, $\mathcal{S}$ respectively. $a_{\mathcal{R}\mathcal{S}}$, b and c are the parameters of the model defined in Eq. (3.1) which are to be estimated from the speech data. Eq. (3.1) shows that $a_{\mathcal{R}\mathcal{S}}$ is a speaker-dependent parameter. We assume b and c to be independent of speaker variability.

In our experiments with the model in Eq. (3.1), the reference speaker was chosen to be the average female of the database. For a given subject speaker, Eq. (3.1) was fitted between the arrays of formant frequencies of the subject and the reference speaker. PnB database consists of 76 speakers (33 males, 28 females and 15 children), each of them contributing two utterances for each of 10 vowels (/AA/, /AE/, /AH/, /AO/, /EH/, /ER/, /IH/, /IY/, /UH/, /UW/). In our analysis, each utterance was considered to be uttered by a different speaker, thus having 152 speakers (66 males, 56 females and 30 children) each uttering 10 vowels. Each of these 10 vowels are characterized by F_1, F_2 and F_3 formant frequencies. An array of frequencies of a given speaker, thus will be of size 30. Eq. (3.1) was fitted between two 30 × 1 frequency vectors.

HiL database *effectively* consists of 98 speakers (37 males, 33 females, 13 boys and 15 girls), each of them uttering only once for each of 12 vowels (/AE/,

/AH/, /AW/, /EH/, /EI/, /ER/, /IH/, /IY/, /OA/, /OO/, /UH/, /UW/). These vowels are also characterized by F_1, F_2 and F_3 formant frequencies. Eq. (3.1) was fitted between two 36×1 frequency vectors for HiL database. The validity of the model was tested for all speakers and the average estimation error energy in fitting the data was calculated to be less than 1.5% of the energy of the corresponding data. Since, Eq. (3.1) has 3 degrees of freedom, each speaker is characterized by 3 parameters. Thus the size of the parameter matrix of the database will be $M \times 3$, where M is the size of the database. Before getting along further, let us ask ourselves few questions. *What is the motivation behind choosing the model in* Eq. (3.1)? *Is this model valid* (in the sense, there may be many models that fit the data better than Eq. (3.1))? Let us answer the first question. The chief motivating factor in choosing the model in Eq. (3.1) was to study whether a "mel-like" frequency-warping function can be obtained from speech data alone. If this is the case, this shows certain connection between the speech production process and the hearing mechanism. This also justifies the use of mel-warp function in speech recognition, not only from psychoacoustic point of view but also from the point of view of speaker normalization.

Taking logarithm on both sides of Eq. (3.1), we have

$$\log(F_{\mathcal{R}}) = \log(a_{\mathcal{R}\mathcal{S}}) + c \log\left(1 + \frac{F_{\mathcal{S}}}{b}\right) \tag{3.2}$$

Define

$$\eta_{\mathcal{S}} = \log(F_{\mathcal{R}}) - \log(a_{\mathcal{R}\mathcal{S}}) \tag{3.3}$$

Thus from Eq. (3.3) we have

$$\eta_{\mathcal{S}} = c \log\left(1 + \frac{F_{\mathcal{S}}}{b}\right) \tag{3.4}$$

where $\eta_{\mathcal{S}}$ represents the formant frequencies of speaker, $\mathcal{S}$ in the warped domain. Similarly, for speaker $\mathcal{Q}$, the warped formant frequencies are given as

$$\eta_{\mathcal{Q}} = \log(F_{\mathcal{R}}) - \log(a_{\mathcal{R}\mathcal{Q}}) = c \log\left(1 + \frac{F_{\mathcal{Q}}}{b}\right) \tag{3.5}$$

Hence, based on Eq. (3.4) and Eq. (3.5), the frequency-warping function to do speaker normalization is given by

$$\eta = c \log\left(1 + \frac{f}{b}\right) \tag{3.6}$$

Database	b	σ_b	c	σ_c
PnB	0.7710	0.3362	0.9756	0.0575
HiL	0.7369	0.2700	0.9761	0.0448

Table 3.1: Estimates of parameters b and c for model based normalization.
The estimates of b *and* c *were calculated by fitting* Eq. (3.1) *for all the speakers of the database with average female as the reference.* σ_b *and* σ_c *are the standard deviations of* b *and* c *respectively.*

Eq. (3.3) and Eq. (3.5) show that in the warped domain η, the speakers are shifted versions of each other, the shift factor being speaker specific. Since the magnitude of Fourier transform is shift-invariant, the features derived are thus shift-invariant in the warped domain and thus resulting in speaker normalization. It is interesting to note that Eq. (3.6) is of the form functionally similar to mel-warp function, suggested by Shaughnessy [17] formula as

$$\eta_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \qquad (3.7)$$

The closeness of Eq. (3.6) to Eq. (3.7) is to be verified by fixing b and c, thus making them speaker-independent. Also, the model with 3 degrees of freedom is very difficult to implement on the recognizer as the parameters are to be estimated on a 3-D mesh, which is terribly cumbersome. These were the two problems that motivated or rather restricted us to reduce the dimensionality of Eq. (3.1) to one, by fixing b and c. The histogram was used in studying the distributions of b and c. Finally, b and c values were fixed to the mean values of their respective distributions.

Table 3.1 shows the estimates of b and c for PnB and HiL databases. The standard deviations, σ_b and σ_c of the parameters b and c respectively, shown in Table 3.1 confirm our assumption that b and c are speaker-independent. Hence the normalization scheme involves using Eq. (3.1) where b and c are chosen from Table 3.1 for appropriate databases. This can be implemented easily on a recognizer, as the parameter search will be over 1-D mesh. The speaker-dependent $a_{\mathcal{RS}}$ can be computed from least squares fit between the formant frequencies of the reference and the subject speaker. This step can be considered to be equivalent to the estimation of k, as discussed in Chapter 2. Figure 3.1 shows the histograms of $a_{\mathcal{RS}}$ for PnB and

HiL databases for male, female and child speakers. It is interesting to note that $a_{\mathcal{RS}}$ and α (in Eq. (2.1)) are inversely related to each other. An average female being the reference, a male subject will have $\alpha < 1$ and a child subject will have $\alpha > 1$. Figure 3.1 shows that male subjects have $a_{\mathcal{RS}} > \rho$ and a child speakers have $a_{\mathcal{RS}} < \rho$ with respect to an average female speaker as reference where ρ is some threshold dependent on the database. The trend in the estimates of $a_{\mathcal{RS}}$ across the genders shows the existence of gender separability. Though, there exist few outliers resulting in crossover of the clusters, the number of such speakers are very less compared to the size of the database. Thus, the model in Eq. (3.1) is valid in the sense of gender separation. The warping function is given as

$$W(f) = \begin{cases} 0.9756 \log \left(1 + \frac{f}{0.7710}\right) & \text{for PnB database,} \\ 0.9761 \log \left(1 + \frac{f}{0.7369}\right) & \text{for HiL database.} \end{cases} \tag{3.8}$$

Having answered our first question, let us try to answer the second one. *Is the model in* Eq. (3.1) *valid? If so, how good it is?*

3.3   Model Validity

To answer the second question, we made a comprehensive study of the relation between speakers, by finding out different models apart from Eq. (3.1) that normalize the speakers. The analysis was carried out on the formant data of speech collected from 14 average speakers for both PnB and HiL databases. 5 average male, 5 average female and 4 average child speakers were obtained for both the databases. Hence the size of the dataset that was considered is reduced to 14 speakers. In this dataset, for a given reference speaker, a subject speaker can be chosen in $\binom{14}{1}$[†] ways (with the repetitions allowed). This results in $\binom{14}{1} \times \binom{14}{1} = 196$ different combinations of the subject and reference speakers. TableCurve2D curve-fitting package was used to fit simple models to each of 196 combinations. This was mainly carried out to search for the best simple fit relationships satisfying all the combinations. The best 20 simple models with least fitting errors were considered for each of 196 combinations. A subjective measure was developed to find the better models out of this whole lot (a total of $196 \times 20 = 3920$, including the repetition of models across the combinations).

[†] $\binom{n}{k} = \frac{n!}{(n-k)!k!}$ is the binomial coefficient
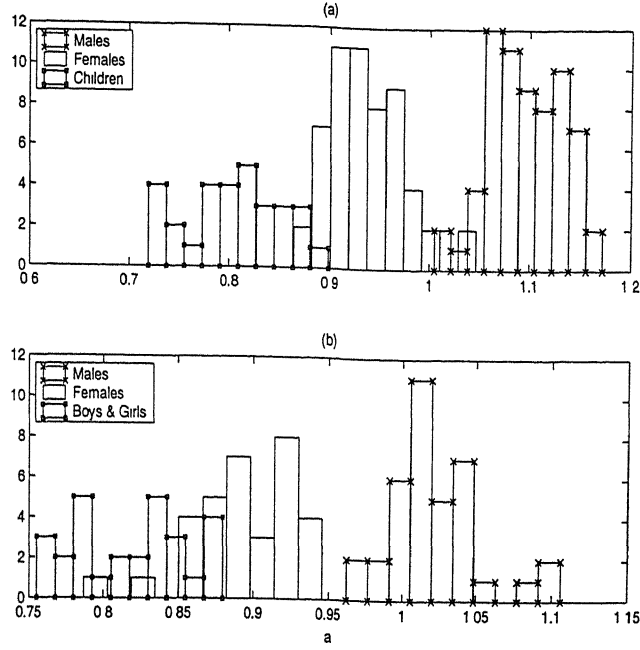
Figure 3.1: Histogram of the speaker dependent parameter, a in model based normalization.

Figure shows the histogram of $a_{\mathcal{RS}}$, *(Eq. (3.1)), plotted for all the speakers in (a) PnB database and (b) HiL database. The histograms clearly show 3 clusters depicting the gender separability. Boy and Girl speakers of HiL database were considered jointly as child speakers.*

The following procedure was developed to find the best models that satisfy all the combinations.

1. Rank the model-list obtained for each combination (The model-list will be arranged in the descending order of accuracy of fit or ascending order of fitting error).

2. Define $N = [N_1 \ N_2 \ \cdots \ N_L]$. N_j (j^{th} element of N), represents the number of occurrences of jth model (in the list of 3920 models). L represents the number of distinct models.

3. Define a measure,

$$Q_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{R_{ij}(100 - A_{ij})}{100}, \ 1 \le j \le L \tag{3.9}$$

R_{ij} and A_{ij} represent the rank and the accuracy (in %) of the jth model in i^{th} combination. Q_j is the jth element of the vector $\mathbf{Q}$ where $\mathbf{Q} = [Q_1 \ Q_2 \ \cdots \ Q_L]$.

4. Calculate

$$\ell = \arg \min_{1 \le j \le L} \mathbf{Q} \tag{3.10}$$

5. ℓ^{th} model is the model (out of L models) that best satisfies/fits all the 196 combinations.

It is evident from Eq. (3.9) that $(100 - A_{ij})$ is the percentage fitting error for j^{th} model of ith combination. Hence, the product $R_{ij}(100 - A_{ij})$ should be small for the models that fit the data better. Table 3.2 shows that Eq. (3.1) is the model that best fits the data for both PnB and HiL databases. This indeed answers our second question. But one should be aware of the *catch* here. The higher-order models fit the data better than the lower-order models. Since, we were interested in one-parameter models, we studied the validity of one-parameter models. The multi-parameter models in Table 3.2 were reduced to one-parameter models by fixing the parameters with less variance. Only the 10 best models in Table 3.2 were considered while reducing the multi-parameter models to one-parameter models. Table 3.3 shows one-parameter models for PnB and HiL databases along with the fitting errors calculated using Eq. (3.11). Suppose $F_{\mathcal{R}} = \hbar(a_{\mathcal{RS}}, F_{\mathcal{S}})$ is the model, the curve-fitting error was estimated as

$$e = \frac{1}{M} \sum_{i=1}^{M} |F_{\mathcal{R}} - \hbar(a_{\mathcal{RS}}, F_{\mathcal{S}})|^2 \tag{3.11}$$

where M is the number of speakers in the database. Table 3.3 substantiates the validity of Eq. (3.1).

Rank	Model Equations	
	PnB	HiL
1	$y = a \log \left(1 + \frac{x}{b}\right)^c$	$y = a \log \left(1 + \frac{x}{b}\right)^c$
2	$y = a + bx^c$	$y = a + be^{-x/c}$
3	$y = ax^b$	$y = ax^b$
4	$y = a + be^{-x/c}$	$y^{0.5} = a + bx^{0.5}$
5	$y^{0.5} = a + bx^{0.5}$	$\log(y) = a + b \log(x)$
6	$y = a + bx$	$y = a + bx$
7	$\log(y) = a + b \log(x)$	$y = a + bx^c$
8	$y^{-1} = a + bx^{-1}$	$y^{-1} = a + bx^{-1}$
9	$y^2 = a + bx^2$	$y^2 = a + bx^2$
10	$y = a + \frac{bx}{\log(x)}$	$y = a + \frac{bx}{\log(x)}$
11	$y^{0.5} = a + bx^{0.5} \log(x)$	$y^{0.5} = a + bx^{0.5} \log(x)$
12	$y^2 = a + bx^2 \log(x)$	$y = a + bx \log(x)$
13	$y = a + bx \log(x)$	$y^2 = a + bx^2 \log(x)$
14	$\log(y) = a + b \left(\log(x)\right)^2$	$\log(y) = a + b \left(\log(x)\right)^2$
15	$y^{-1} = a + bx^{-1} \log(x)$	$y^{-1} = a + bx^{-1} \log(x)$

Table 3.2: Best simple curvefits for vowel data.

The equations with smaller rank are the models that best fit the data. It is interesting to note the similarity in the models for PnB and HiL databases. Eq. (3.1) is the model that fits the data best for both PnB and HiL databases. This answers the validity in choosing Eq. (3.1).

3.4 Comparison of Model Based Frequency Warping Function and Mel-Warp Function

As mentioned earlier, one of the motivating factors in choosing the model in Eq. (3.1) is the functional similarity of Eq. (3.6) with Eq. (3.7). We made a comprehensive study to verify the existence of *mel-like* warping function obtained only from speech data. Since the model in Eq. (3.1) is non-linear, the error performance surface may have many minimas. Let us define $T = (a_{\mathcal{RS}}, b, c)$. The estimation of parameters defined by T requires the initial estimate of the parameters along with the search

Rank	PnB		HiL	
	Model	e	Model	e
1	$y = a\log\left(1 + \frac{x}{0.77}\right)^{0.98}$	1.98E4	$y = a\log\left(1 + \frac{x}{0.74}\right)^{0.98}$	1.40E4
2	$y = a + 2.64x^{0.93}$	3.68E4	$y = a + 1.92x^{0.96}$	2.28E4
3	$y = ax^{0.98}$	1.98E4	$y = ax^{0.98}$	1.40E4
4	$y = a + e^{-x}$	9.92E5	$y = a + e^{-x}$	9.86E5
5	$y^{0.5} = a + x^{0.5}$	2.40E4	$y^{0.5} = a + x^{0.5}$	1.55E4
6	$y = a + 1.04x$	4.02E4	$y = a + 1.02x$	2.38E4
7	$\log(y) = a + 0.98\log(x)$	1.98E4	$\log(y) = a + 0.98\log(x)$	1.40E4
8	$y^{-1} = bx^{-1}$	2.02E4	$y^{-1} = bx^{-1}$	1.43E4
9	$y^2 = a + 1.11x^2$	9.18E4	$y^2 = a + 1.07x^2$	6.31E4
10	$y = a + \frac{8.76x}{\log(x)}$	3.52E4	$y = a + \frac{8.64x}{\log(x)}$	2.17E4

Table 3.3: Best one parameter models for vowel data.
Here, the rank does not signify the quality of fit. The one- parameter version of Eq. (3.1) is the best model along with few other models that fits the data with least error for both PnB and HiL databases. This answers the validity in choosing Eq. (3.1).

range. The values of b and c in Table 3.1 were obtained by choosing the initial estimates to be 1 and range of search to be $\Re$, i.e. $(-\infty, \infty)$. The solution for **T** obtained with these initial estimates may not be the global minima. It is important to note that b is the parameter that actually determines the shape of the warping function. c is just a scaling factor that is hardly of any importance. We developed the following procedure to find the initial estimate of b (over the region of interest) that gives the minimum error according to Eq. (3.11). We mainly carried out this procedure to check whether an initial estimate of b around 500 to 1000, would provide the least residue than the other initial estimates of b.

1. Define $\mathbf{B} = [b_1 \ b_2 \ \cdots \ b_I]$, where b_j is the jth initial estimate of b and I is the number of different initial estimates. Choose the initial estimates of a and c to be 1.

2. Fit the model defined in Eq. (3.1) for a given speaker with different initial estimates $\mathbf{T}_{initial} = (1, b_j, 1)$ and $a_{\mathcal{RS}}, b, c \in \Re$, $1 \leq j \leq I$.
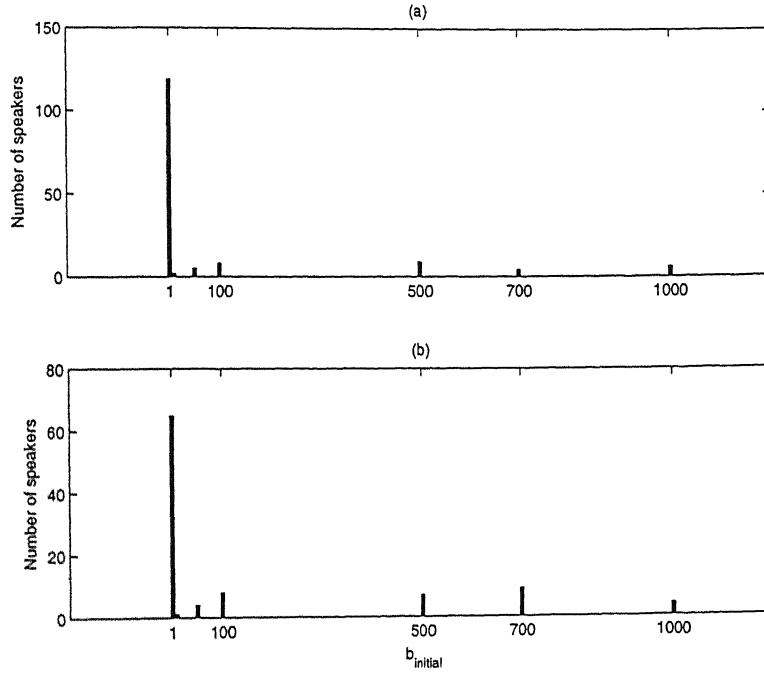
Figure 3.2: Histogram of the initial estimates of b for model based normalization. *Figure shows the histogram of the initial estimates of* b *(Eq. (3.1)) for (a) PnB database and (b) HiL database. In our experiments.* B = [1 10 50 100 500 700 1000]. *It is clear that* $b_{initial}$ = 1 *gives minimum* e *(Eq. (3.12)) for both PnB and HiL databases.*

3. Calculate the residue as

$$e_j = \sum \left| F_{\mathcal{R}} - a_{\mathcal{RS}} \left(1 + \frac{F_{\mathcal{S}}}{b}\right)^c \right|^2 \tag{3.12}$$

4. Determine the initial estimate of b that minimizes the residual defined in Eq. (3.12) as $\imath^{th}$ element of **B**, where

$$\imath = \arg \min_{1 \leq j \leq \mathrm{I}} e \tag{3.13}$$

and $e = [e_1 \ e_2 \ \cdots \ e_{\mathrm{I}}]$.

5. Repeat the steps 2 to 4 for all speakers in the database.

Figure 3.2 shows the histograms of the initial estimates of b for PnB and HiL databases. It is clear from Figure 3.2 that $b_{initial}$ = 1 has the maximum of number
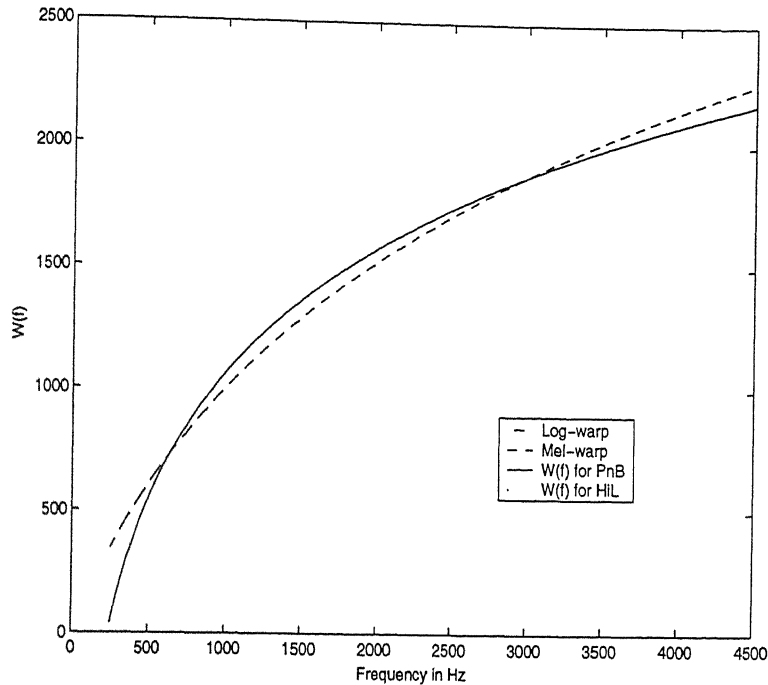
Figure 3.3: Comparison of warping function, $W(f)$ derived from model based normalization with log-warp and mel-warp functions.

Figure shows the warping functions derived using MBN method (Eq. (3.8)) for PnB and HiL databases, along with log-warp and mel-warp functions. Here, log-warp function corresponds to the model defined in Eq. (3.14).

of speakers for whom e (Eq. (3.12)) is minimum compared to other initial estimates of b. This analysis thus supports the warping functions defined in Eq. (3.8), which were derived with $T_{initial} = (1, 1; 1)$. Hence the warping functions derived from Eq. (3.1) are indeed reliable. Figure 3.3 shows $W(f)$ for PnB and HiL databases along with log-warp and mel-warp functions. It shows that $W(f)$ is very close to log-warp function than to mel-warp function. Hence our motivation to verify the similarity between the speech-derived warping function and mel-warp function resulted in a *more* log-like function, which in turn refers to uniform scaling between the speakers. Now, here arises the *contradiction*. In Chapter 2, we showed that the warping function derived from the speech data is a compromise between the log-warp and mel-warp functions. But this analysis of speaker relationships has resulted in a log-like function. *How does this difference come about?*

3.5 Comparison of Model Based Frequency Warping Function and Log-Warp Function

Inorder to convince ourselves with the result explained in the previous section, we experimented with a model different from Eq. (3.1), given by

$$F_\mathcal{R} = a_{\mathcal{R}\mathcal{S}} F_\mathcal{S} \qquad (3.14)$$

which is uniform scaling of the formants of two speakers, $\mathcal{R}$ and $\mathcal{S}$, the scaling factor, $a_{\mathcal{R}\mathcal{S}}$ being only speaker dependent and is independent of frequency. Applying logarithm to both sides of Eq. (3.14), we have

$$\log(F_\mathcal{R}) = \log(a_{\mathcal{R}\mathcal{S}}) + \log(F_\mathcal{S}) \qquad (3.15)$$

Hence the warping function is given by

$$\mathsf{W}(f) = \log(f) \qquad (3.16)$$

which is log-warp function.

A comparative study on the performance of various normalization procedures (discussed in Chapter 2) was carried out in a qualitative and quantitative sense to study the effectiveness of normalization. The results of these experiments, explained in Chapter 4 and Chapter 6 shows that the model in Eq. (3.14) does better normalization than simple linear scaling of Nordström & Lindblom, Fant's non-uniform normalization and frequency dependent scaling methods. This shows that uniform scaling is better than other normalization methods. *This is very interesting.* It is important to note that the warping function derived from FDS method is actually a discrete one. The discretization of the warping function has resulted in fine modelling of the non-linearity of the scaling factor in each frequency region, thus deviating from uniform scaling. But, the warping function derived from Eq. (3.1) is a continuous function of frequency, the parameters being obtained by fitting Eq. (3.1) to the formant data of subject and reference speakers. Hence, on an average, this produces a uniform scale-like relationship. The performance difference (discussed in Chapter 4) between the linear scaling method of Nordström & Lindblom and uniform scaling obtained by MBN is quite interesting, the latter outperforming the former. The only difference between these two methods is the

parameters that are involved in the computation of scale factors. In Nordström & Lindblom method, only *average F_3* in *open vowels* was considered in computing the scale factor, whereas in MBN method, all the formants, (F_1, F_2, F_3) were involved in the computation of a_{RS}. On the whole, MBN provides a gross linear-scale relation among the speakers.

3.6 Summary

A model based vowel normalization procedure, motivated by the idea to study the relationship between the formant frequencies of speakers was presented. The warping function derived from the assumed model was compared with log-warp and mel-warp functions. Model based normalization has resulted in a linear scale relationship among the speakers, which is a gross approximation to the non-linearity present in their actual relationship.

Chapter 4

Comparison of Vowel Normalization Methods Using Separability Measures

In the previous chapters, we discussed different methods of vowel normalization, each of them aimed at reducing the speaker dependence. The warping functions were derived with an aim to apply these normalization methods to continuous spectral patterns. The criterion for the degree of success of these normalization procedures might be that they should maximally reduce the variance within each group of vowels when spoken by different speakers, while maintaining the separation between such groups. In this chapter, we compare these different vowel normalization methods by defining various measures, both in qualitative and quantitative sense.

4.1 Residual Variance

One of the measures [5] used by Fant to find the efficacy of the non-uniform normalization scheme is the percentage of variance remaining after non-uniform normalization when compared to the uniform normalization scheme of Nordström & Lindblom [1]. The variance in each of the three formants, F_1, F_2, F_3 after normalization is given by

$$V_n = \sum_{subject} \sum_{vowel} |k_{n,observed} - k_{n,predicted}|^2, \quad n = 1, 2, 3 \qquad (4.1)$$

where $k_{n,observed}$ is calculated using the actual value of the nth formant of each vowel of the subject and the reference speaker (i.e. average female). $k_{n,predicted}$ is the predicted value of scale factor for the nth formant of each vowel of the subject using a given vowel normalization scheme. The percentage residual variance after non-uniform normalization compared to uniform normalization of Nordström & Lindblom for the nth formant is defined as

$$R_n = \frac{V_{n,non-uniform}}{V_{n,uniform}} \times 100 \qquad (4.2)$$

In our experiments with vowel normalization methods, we computed R_n for Fant's normalization method, frequency dependent scaling method and model based normalization method. Table 4.1 shows the performance of different normalization schemes against the uniform normalization method. It is clear from Table 4.1 that the performance of FDS and MBN methods is comparable to Fant's method even though they assume no *a priori* information about the vowel category and formant number unlike Fant's method. Further, for HiL data, it can be seen that MBN outperforms Fant's method especially for children.

4.2   F-Ratio

Since discriminability between vowel clusters is as important as reduction of variance within any given vowel cluster, a good measure for the usefulness of the normalization schemes would be F-ratio [10, 21]. In discriminant analysis, within-class and between-class scatter matrices are used to formulate criteria of class separability. In deriving F-ratio, one of the separability measures, let M_i and C_i denote the mean formant (F_1, F_2, F_3) vector and its covariance matrix respectively, of the ith vowel class. An equal probability of vowel classes is assumed. Let $M_0 = \frac{1}{I} \sum_{i=1}^{I} M_i$, where I denotes the number of vowel classes being compared. Then, the within-class, S_w and between-class, S_b scatter matrices, are computed by

$$S_w = \frac{1}{I} \sum_{i=1}^{I} C_i \qquad (4.3)$$

$$S_b = \frac{1}{I} \sum_{i=1}^{I} (M_i - M_0)(M_i - M_0)^T \qquad (4.4)$$

Residual Variance (%)		PnB			HiL		
		Ad. & Ch.	Ad.	Ch.	Ad. & Ch.	Ad.	Ch.
Fant	R_1	90	80	108	103	75	151
	R_2	80	78	84	89	78	97
	R_3	93	92	96	78	83	74
FDS	R_1	88	86	91	101	84	130
	R_2	78	82	72	81	81	81
	R_3	100	97	106	82	86	79
MBN−1	R_1	93	96	85	80	77	84
	R_2	72	79	62	79	74	83
	R_3	84	84	84	73	78	69
MBN−2	R_1	88	90	85	87	80	98
	R_2	79	79	65	78	76	80
	R_3	88	88	89	76	81	71

Table 4.1: Residual variance after normalization.

Percentage variance remaining after different non-uniform normalization methods when compared to uniform normalization of Nordström & Lindblom, for the three formants. Here Ad. *stands for adult speakers and* Ch. *stands for child speakers. Average female of the respective databases was considered as the reference speaker in computing residual variance.* MBN−1 *refers to the model* $F_{\mathcal{R}} = a_{\mathcal{R}S}\left(1 + \frac{F_S}{b}\right)^c$ *and* MBN−2 *refers to the model* $F_{\mathcal{R}} = a_{\mathcal{R}S}F_S$ *defined in* Chapter 3.

where T represents matrix transposition. The separability criterion is then given by,

$$J = trace\{(\mathbf{S}_b + \mathbf{S}_w)^{-1}\mathbf{S}_b\} \tag{4.5}$$

The vowel cluster discriminability in terms of F-ratio, J, for unnormalized (unwarped), uniform normalization, Fant's non-uniform normalization, FDS and MBN methods are shown in Table 4.2. It is clear from Table 4.2 that, MBN does the best normalization followed by FDS method, both of them being context and formant independent unlike Fant's method. In Eq. (4.5), as separability improves, J should approach the ideal value of 3.

F-Ratio (J)	PnB			HiL		
	Ad. & Ch.	Ad.	Ch.	Ad. & Ch.	Ad.	Ch.
Un-normalized	2.01	2.21	2.31	2.13	2.28	2.31
Nordström & Lindblom	2.42	2.45	2.43	2.47	2.56	2.37
Fant's non-uniform	2.49	2.52	2.41	2.52	2.63	2.40
FDS	2.47	2.50	2.47	2.53	2.61	2.44
MBN−1	2.49	2.51	2.50	2.56	2.62	2.46
MBN−2	2.49	2.50	2.50	2.62	2.50	2.46

Table 4.2: Vowel cluster discriminability in terms of F-Ratio.
Performance of various vowel normalization schemes based on F-ratio measure, applied on PnB and HiL databases. Ad., Ch., MBN−1, MBN−2 *are the same as explained in* Table 4.1.

4.3 Scatter Plots

The $F_1 - F_2$ scatter plots for PnB and HiL databases are shown in Figure 4.1 and Figure 4.2. Larger the separation between the clusters and smaller the spread of the cluster, better is the normalization. Scatter plots in Figure 4.1 and Figure 4.2 provide a visual measure showing that MBN and FDS methods provide better normalization than Fant and Nordström & Lindblom methods.

4.4 Summary

Different vowel normalization methods were compared both in both qualitative and quantitative sense. The efficacy of vowel normalization methods has to be judged from the size of the vowel clusters and the separation between the clusters after normalization. The quantitative measures like residual variance and F-ratio along with the qualitative measures like scatter plots were used to compare different normalization schemes. Model based normalization and frequency dependent scaling methods perform better than Fant and uniform scaling methods with respect to all the aforesaid measures.

Figure 4.1: Scatter plots of $F_1 - F_2$ for 10 vowels from Peterson & Barney database. *Figure shows the scatter plots of $F_1 - F_2$ for 10 Vowels from Peterson & Barney database with and without normalization. MBN—1, MBN—2 are the same as explained in Table 4.1. As seen in the figure MBN—1 and MBN—2 followed by FDS method provides good separability among vowel clusters.*
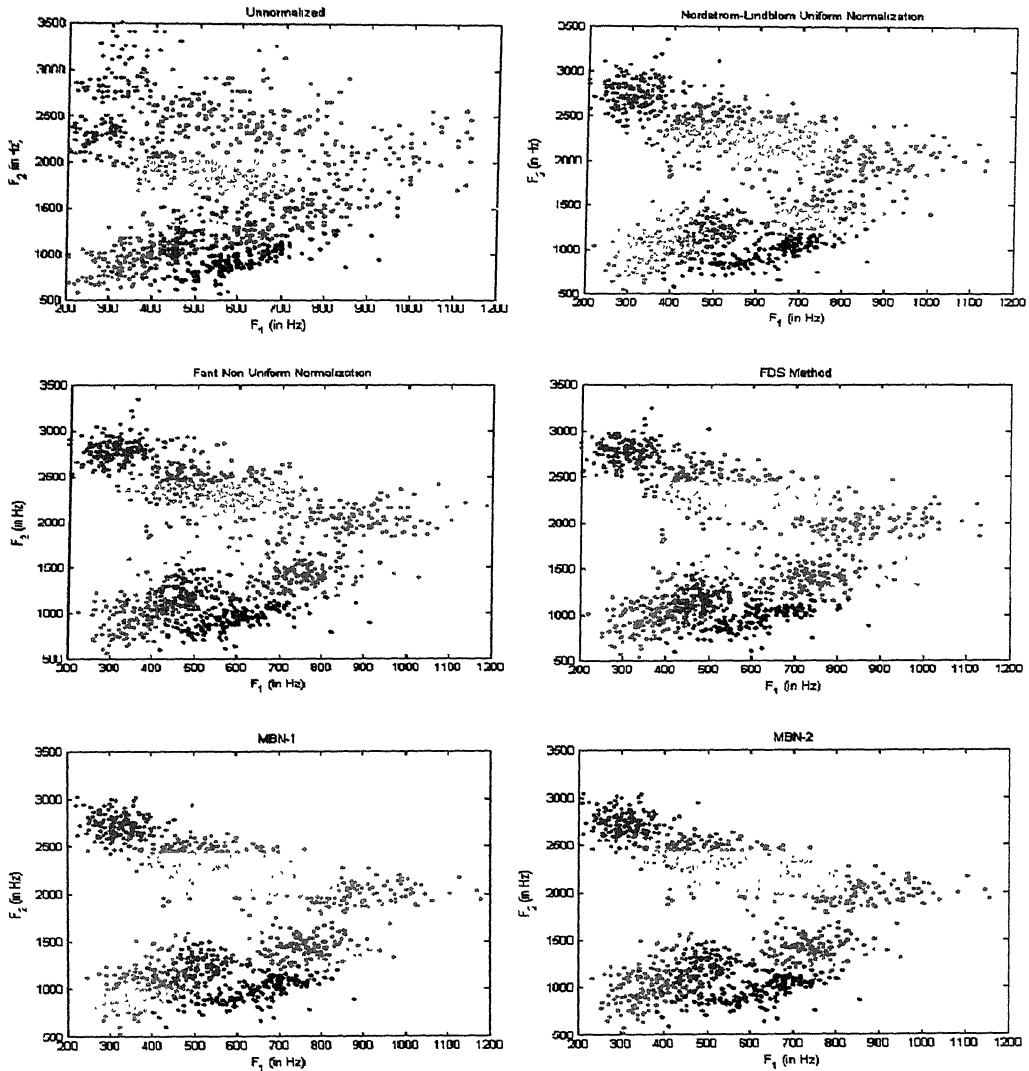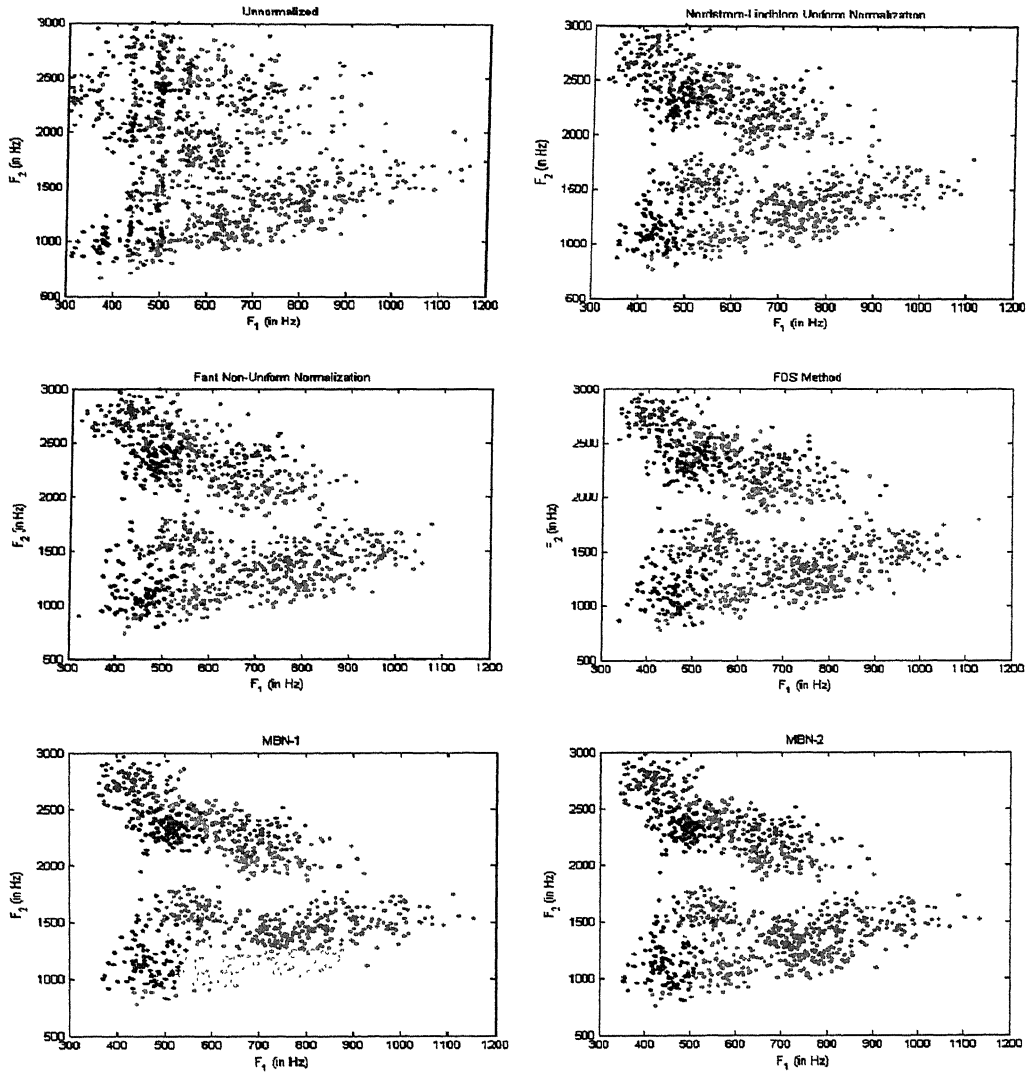
Figure 4.2: Scatter plots of $F_1 - F_2$ for 12 Vowels from Hillenbrand *et al.* database. *Figure shows the scatter plots of $F_1 - F_2$ for 12 Vowels from Hillenbrand et al. database with and without normalization. MBN−1, MBN−2 are the same as explained in* Table 4.1. *As seen in the figure MBN−1 and MBN−2 followed by FDS method provides good separability among vowel clusters.*

Chapter 5

Estimation of Frequency-Warping Function Using Vowel Data

The concept of differences in the vocal tract length in speakers leading to the major source of variability in speech is well established. This has lead to the "scale relationship" between the speakers. The motivation of speaker normalization has resulted in the application of scale invariant transforms (viz. Scale Transform [10, 22], Fourier–Mellin Transform [23]) in deriving the speech features. The basic idea is to warp a pair of mutually scaled spectra such that in the warped domain they are shifted versions of one another. By taking the magnitude of Fourier transform of these shifted versions, we get identical spectras in the warped domain.

5.1   Scale Transform

Briefly, the scale transform of a function, $X(f)$ is given by,

$$D_X(c) = \int_0^\infty X(f) \, \frac{e^{-j2\pi c \ln f}}{\sqrt{f}} \, df \qquad (5.1)$$

and inversely,

$$X(f) = \int_{-\infty}^\infty D_X(c) \, \frac{e^{j2\pi c \ln f}}{\sqrt{f}} \, dc \quad \forall \ f \geq 0 \qquad (5.2)$$

A basic property of the scale transform is that the magnitude of the scale transform of a function, $X(f)$ and its normalized scaled version, $\sqrt{\alpha}X(\alpha f)$, are equal (note that

$0 < \alpha < 1$ corresponds to dilation, while $1 < \alpha < \infty$ corresponds to compression), since

$$D_{\mathsf{X}}^{\alpha}(c) = \int_{0}^{\infty} \sqrt{\alpha}\mathsf{X}(\alpha f) \frac{e^{-j2\pi c \ln f}}{\sqrt{f}} \, df \tag{5.3}$$

$$= e^{j2\pi c \ln \alpha} D_{\mathsf{X}}(c) \tag{5.4}$$

Eq. (5.4) shows that the scale transform of $\sqrt{\alpha}\mathsf{X}(\alpha f)$ is same as that of $\mathsf{X}(f)$ except for a linear phase, which disappears by taking magnitude on both sides of Eq. (5.4), i.e.,

$$|D_{\mathsf{X}}^{\alpha}(c)| = |D_{\mathsf{X}}(c)| \tag{5.5}$$

Thus, considering two speakers who are scaled versions of one another, α being their characteristic scale factor, Eq. (5.5) shows that in the scale transform domain, both the speakers look alike, as the speaker dependent term that appears in phase is nullified by taking magnitude. Thus, there is no need to explicitly calculate the speaker specific scaling constant. The scale transform may also be calculated as the fourier transform of the function $\mathsf{X}(e^f)e^{\frac{f}{2}}$ i.e.,

$$D_{\mathsf{X}}(c) = \int_{-\infty}^{\infty} \mathsf{X}(e^f)e^{\frac{f}{2}} \, e^{-j2\pi c f} \, df \tag{5.6}$$

It is to be noted that as a result of log-warping, i.e., forming $\mathsf{X}(e^f)$, the speaker specific scale constant, α, is purely a function of translation parameter in the log-warped domain.

5.2   Frequency Warping Function

Consider two speakers $\mathcal{A}$ and $\mathcal{B}$ related by

$$\mathsf{S}_{\mathcal{A}}(f) = \mathsf{S}_{\mathcal{B}}(\alpha_{\mathcal{A}\mathcal{B}}f) \tag{5.7}$$

where $\mathsf{S}(.)$ denotes the spectral envelopes and $\alpha_{\mathcal{A}\mathcal{B}}$ is the scale factor of the subject speaker $\mathcal{B}$, with respect to the reference speaker, $\mathcal{A}$. This would be the case if uniform scaling was true. Consider the warping function $f = e^v$ which is applied to

all speakers. This is log-warping of the frequency axis. In the log-warped domain, the scaling factor α_{AB} appears as a translation factor.

$$
\begin{aligned}
S_a(v) &= S_A(f = e^v) = S_B(\alpha_{AB}e^v) \\
&= S_B(e^{v + \ln \alpha_{AB}}) = S_b(v + \ln \alpha_{AB})
\end{aligned}
\tag{5.8}
$$

Note the use of lower-case subscripts that denote the spectras or functions in the warped domain or v-domain. Thus, in the log-warped domain, the warped spectras are shifted versions of one another. The magnitude of their Fourier transform leads to scale invariance.

$$
|F(S_A(v))| = |F(S_B(v + \ln \alpha_{AB}))|
$$

where $F(.)$ represents the fourier transform operator. Here exponential sampling denotes linear scaling of the frequency axis, which is realized as equal sampling in log-domain. Figure 2.1 shows that the scale factor, α_{AB} is indeed formant dependent and context dependent, resulting in non-uniform scaling. In such a case, the relation between spectral envelopes of two speakers can be modelled as

$$
S_A(f) = S_B(\alpha_{AB}(f)f)
\tag{5.9}
$$

where $\alpha_{AB}(f)$ is a frequency-dependent, non-uniform scaling factor. Analogous to uniform scaling, our goal is to find a transformation, $f = z(v)$ such that

$$
\begin{aligned}
S_a(v) &= S_A(f = z(v)) \\
&= S_B(\alpha_{AB}(f)f) = S_b(v + \varsigma_{AB})
\end{aligned}
\tag{5.10}
$$

where ς_{AB} is dependent only on the speakers A and B, and is independent of frequency. Now, our aim is to find the function $z(.)$, which warps the spectras, thus making them shifted versions in the warped domain. Since finding the exact form of $z(.)$ is difficult, we discretized the computation of the warping function [7, 24].

5.3 Numerical Computation of the Warping Function

5.3.1 Discrete-Implementation of Warping Function

We now obtain a relationship between f and v at a discrete set of frequencies. Let us divide the frequency axis into N logarithmically equi-spaced regions. In each region,

let us assume that the spectral envelopes of any two speakers are scaled versions of one another. So, for i^{th} frequency region, $f \in [L_i, U_i)$, we have

$$S_{\mathcal{A}}(f) = S_{\mathcal{B}}(\alpha_{AB}^{(i)} f) \ , \ L_i \leq f < U_i \tag{5.11}$$

where $\alpha_{AB}^{(i)}$ is the scale factor for i^{th} frequency region, L_i and U_i are the lower and upper frequency boundaries of ith region respectively and $1 \leq i \leq N$. Define

$$\alpha_{AB}^{(i)} = \alpha_{AB}^{\beta_i} \tag{5.12}$$

which assumes that the frequency dependency is present in the parameter β and α_{AB} is only speaker dependent (independent of frequency). We need to compute $S_b(v = \log(f))$ for $v \in [\log(L_i), \log(U_i))$. Let us discretize the computation of $S_b(v)$ at M_i equally spaced intervals in the region $\log(L_i)$ to $\log(U_i)$. Let

$$\Delta v_i = \frac{\log(U_i) - \log(L_i)}{M_i} \tag{5.13}$$

Then the uniformly spaced samples in the ith frequency region in v-domain are $S_b(m_i \Delta v_i + \log(L_i))$ for $m_i = 0, 1, \cdots, (M_i - 1)$. Uniformly sampling $S_a(v)$ at Δv_i spacing in the ith frequency region results in

$$S_a(m_i \Delta v_i + \log(L_i)) = S_b\left(m_i \Delta v_i + \log(\alpha_{AB}^{\beta_i}) + \log(L_i)\right) \tag{5.14}$$

Eq. (5.14) can be rewritten as

$$S_a(m_i \Delta v_i + \log(L_i)) = S_b\left(\left(m_i + \frac{\beta_i \log(\alpha_{AB})}{\Delta v_i}\right)\Delta v_i + \log(L_i)\right) \tag{5.15}$$

It can be seen that the two functions differ by a translation factor $\frac{\beta_i \log(\alpha_{AB})}{\Delta v_i}$ in the ith frequency region. Since we define the warped envelopes to be translated versions of one another, over the entire range of interest, we require the following condition to be satisfied between any two frequency regions i and j.

$$\frac{\beta_i \log(\alpha_{AB})}{\Delta v_i} = \frac{\beta_j \log(\alpha_{AB})}{\Delta v_j} = \frac{1}{\Delta \lambda} \tag{5.16}$$

where λ is a new domain where the scaled spectras appear as shifted versions of one another. From Eq. (5.13), we have

$$\Delta v_i M_i = \Delta v_j M_j \tag{5.17}$$

as $\log\left(\frac{U_i}{L_i}\right) = \log\left(\frac{U_j}{L_j}\right)$, thus resulting in $\beta_i M_i = \beta_j M_j$. We can therefore choose M_i for different frequency regions i.e., the spacing of samples in v-domain, such that $\beta_i M_i = \beta_j M_j$. The total number of samples, held constant, M_i's are given by

$$\sum_{i=1}^{N} M_i = M_{const} \tag{5.18}$$

$$M_i = \frac{\frac{M_{const}}{\beta_i}}{\sum_{j=1}^{N} \frac{1}{\beta_j}} \tag{5.19}$$

With this choice of M_i's, the non-uniformly spaced samples in v-domain are represented as uniformly spaced samples in λ-domain. Since the scale is arbitrary in λ-domain, we can choose the spacing of samples and origin to some convenient values. Eq. (5.19) shows that the calculation of M_i's depend on β_i's. So, we need to devise a procedure to compute β_i's from the speech data, from which the warping function can be derived easily. This is the situation that was explained in Section 2.4.1. So, given the warping parameters (methods of computation of these parameters may be different), we can numerically compute the discrete warping function. Before examining the method of computation of warping parameters, let us consider the following situation. If there exists a simple linear scaling between two speakers, say $\mathcal{A}$ and $\mathcal{B}$, then $\alpha_{AB}(f) = \alpha_{AB}^{\beta(f)} = \alpha_{AB}$. The scaling factor is only speaker dependent and is independent of frequency. So, $\beta(f) = 1$ or in its discrete form, $\beta_i = 1$, $i = 1, 2, \cdots, N$. The warping function in such a case is given by

$$W(f) = \lambda = v = \log(f) \tag{5.20}$$

Because of the non-linear scaling between the speakers, the scale factor will be both frequency dependent and speaker dependent. In such cases, $\beta(f)$ models the non-linearity in the scale factor. In other words, the non-linearity in i^{th} frequency region is modelled by β_i. Hence the discrete warping function is given by

$$W_i(f) = \lambda = \frac{v}{\beta_i} = \frac{\log(f)}{\beta_i}, \ i = 1, 2, \cdots, N \tag{5.21}$$

5.3.2 Band Edge Problem

The discussion in Section 5.3.1 shows that the sampling rate in v-domain changes abruptly at the band edges. Hence, though the sampling is uniform within a given

band, it is non-uniform over the whole v-domain. The result is the loss of spectral samples at the band edges. To avoid this loss of spectral samples, we carried the following transition band analysis. Eq. (5.15) shows that the warped spectras in i^{th} frequency region are shifted versions of one another, the shift being frequency dependent, as it is a function of β_i. It is the β_i that determines the spacing of samples in v-domain, whose discontinuity at the band edge results in the loss of spectral samples. One way to avoid this problem is to make β_p change gradually to β_{p+1}, $p = 1, 2, \cdots, N - 1$, over a region of K samples, i.e., we need

$$\frac{\beta_p}{\Delta v_p} = \frac{\beta_p + \Delta\beta}{\Delta v_p + \delta_{p,1}} = \cdots = \frac{\beta_p + k\Delta\beta}{\Delta v_p + \delta_{p,k}} = \cdots = \frac{\beta_p + K\Delta\beta}{\Delta v_p + \delta_{p,K}} = \frac{\beta_{p+1}}{\Delta v_{p+1}} \qquad (5.22)$$

where $\Delta\beta$ is a factor which provides a transition in the values of β across the adjacent regions, K is the number of frequency points over which β_p gradually changes to β_{p+1} defined as $K = \left| \frac{\beta_{p+1} - \beta_p}{\Delta\beta} \right|$, $p = 1, 2, \cdots, N - 1$ and $k = 1, 2, \cdots, K$. $\{\delta_{p,k}, k = 1, 2, \cdots, K\}$ are the factors that are to be computed which provide the gradual change in the sampling intervals across two adjacent frequency regions, thus avoiding the loss of spectral samples at the band edges. Hence, given β_p, β_{p+1} and $\Delta\beta$, the factors $\{\delta_{p,k}, k = 1, 2, \cdots, K\}$ can be determined from Eq. (5.22). We consider L points to the either side of the band edge over which β_p changes to β_{p+1}, thus amounting to a total of K points, where L is given as

$$L = \begin{cases} \frac{K-1}{2} & , \text{K is odd} \\ \frac{K}{2} - 1 & , \text{K is even} \end{cases} \qquad (5.23)$$

From Eq. (5.16), it is clear that $\Delta v_i \propto \beta_i$. Hence we have the following cases.

1. $\beta_{p+1} < \beta_p$: $\Delta\beta < 0$ and $\{\delta_{p,k}, k = 1, 2, \cdots, K\}$ forms a decreasing sequence.

2. $\beta_{p+1} = \beta_p$: $\Delta\beta = 0$ and $\{\delta_{p,k} = \delta_{p,k+1}, k = 1, 2, \cdots, K\}$

3. $\beta_{p+1} > \beta_p$: $\Delta\beta > 0$ and $\{\delta_{p,k}, k = 1, 2, \cdots, K\}$ forms an increasing sequence.

It is to note that the above analysis to override the band edge effects may not be the optimum method. In our case, β_p varies linearly within the transition band to β_{p+1}. Different variations can be tried out in the transition of β_p within the transition band. Smaller the value of $\Delta\beta$, the transition from β_p to β_{p+1} will be smooth over

a large number of points. Once v-domain is discretized overriding the band edge effects, the warping function can be computed as explained in Section 5.3.1 using Eq. (5.21) except that the value of β_p in the transition region between p^{th} and $(p+1)^{th}$ frequency regions should be taken as $\beta_i + k\Delta\beta$, where $k = 1, 2, \cdots K$.

5.3.3 Experimental Determination of Warping Parameters

The warping parameters $\alpha_{AB}^{(i)}$ and β_i were computed experimentally from the vowel data of PnB and HiL databases. We had chosen $N = 5$, thus obtaining 5 logarithmically equi-spaced frequency regions. The reason for choosing N to be 5 will be explained later. Table 5.1 shows the frequency regions of interest for PnB and HiL databases. While estimating $\alpha_{AB}^{(i)}$, only two speakers were considered at a time, considering only those pair of formants that lie within the same frequency region. For example, for each pair of speakers, $\mathcal{A}$ and $\mathcal{B}$, we computed the ratio of formants in the ith frequency region as

$$\tau_{AB}^{(i,j,k)} = \frac{F_{\mathcal{B}}^{i,j,k}}{F_{\mathcal{A}}^{i,j,k}} \quad \text{if} \ \ F_{\mathcal{A}}^{i,j,k}, F_{\mathcal{B}}^{i,j,k} \in [L_i, U_i) \tag{5.24}$$

$F_{\mathcal{A}}^{i,j,k}$, $F_{\mathcal{B}}^{i,j,k}$ are the kth formants of the jth vowel of speakers $\mathcal{A}$ and $\mathcal{B}$ respectively and both of them lie in the same i^{th} frequency region. We computed $\tau_{AB}^{(i,j,k)}$ for all pairs of such formants that lie in the ith frequency region and obtained the average scaling factor, $\alpha_{AB}^{(i)}$ as the average of $\tau_{AB}^{(i,j,k)}$ in i^{th} region for a given pair of speakers, $\mathcal{A}$ and $\mathcal{B}$. The estimates of $\alpha_{AB}^{(i)}$ obtained were averaged over to find $\alpha^{(i)}$ representing the frequency dependent scaling factors. β_i's were estimated from the estimates of $\alpha^{(i)}$ as

$$\beta_i = \begin{cases} \frac{\log(\alpha^{(i)})}{\log(\alpha^{(N)})} & \text{for } 1 \leq i \leq N-1, \\ 1 & \text{for } i = N. \end{cases} \tag{5.25}$$

Since the higher formants are mostly affected by the length of the pharyngeal cavity, the uniform scaling holds and hence, we assumed $\beta_N = 1$, thus making $\alpha^{(N)}$ to be the ratio of formants in N^{th} frequency region.

PnB			HiL		
Band (Hz)	β_i	σ_{β_i}	Band (Hz)	β_i	σ_{β_i}
[190,356)	2.13	0.13	[310,524)	1.50	0.03
[356,667)	1.22	0.05	[524,893)	1.55	0.03
[667,1249)	1.51	0.05	[893,1523)	1.46	0.03
[1249,2339)	1.27	0.04	[1523,2598)	1.40	0.02
[2339,4381)	1.00	0.00	[2598,4431)	1.00	0.00

Table 5.1: Average estimates of β_i in 5 logarithmically equi-spaced frequency regions. σ_{β_i} *denotes the standard deviation of* β_i *for* i^{th} *frequency band. Here,* $1 \leq i \leq 5$

Database	$W(f)$
PnB	$-1203.48 + 47.57 \left(\log(f)\right)^2$
HiL	$-1323.47 + 49.70 \left(\log(f)\right)^2$

Table 5.2: Closed form equations for discrete warping function, $W_i(f)$.
$W(f)$ *is the closed form equation for the discrete warping function,* $W_i(f)$. *The curve-fits were obtained by using TableCurve2D package*

5.4   Experiments and Results

The experiments were carried out by overriding the band edge problems to obtain the discrete warping function. Table 5.1 shows the estimates of β_i along with their standard deviations for both PnB and HiL databases, obtained by averaging over all speakers. Figure 5.1 shows the discrete warping function, $W_i(f)$ obtained with and without transition band analysis for PnB and HiL databases. Though, the warping functions look similar, practically, it is important to override the band edge problem. We chose $|\Delta\beta| = 0.0275$ for PnB database and $|\Delta\beta| = 0.010$ for HiL database. Depending on the sign of $(\beta_{p+1} - \beta_p)$ being +ve or −ve, $\Delta\beta$ was chosen to be +ve or −ve for the pth band. The reason for choosing N to be 5 is to model the non-linearity in a better way. Smaller the value of N, more coarsely will be the modelling of the non-linearity. Larger values of N results in finer modelling of the non-linearity. But, large values of N results in less data available for the estimation of $\alpha^{(i)}$ thus questioning its reliability. Hence, a trade-off was to be made between the finer
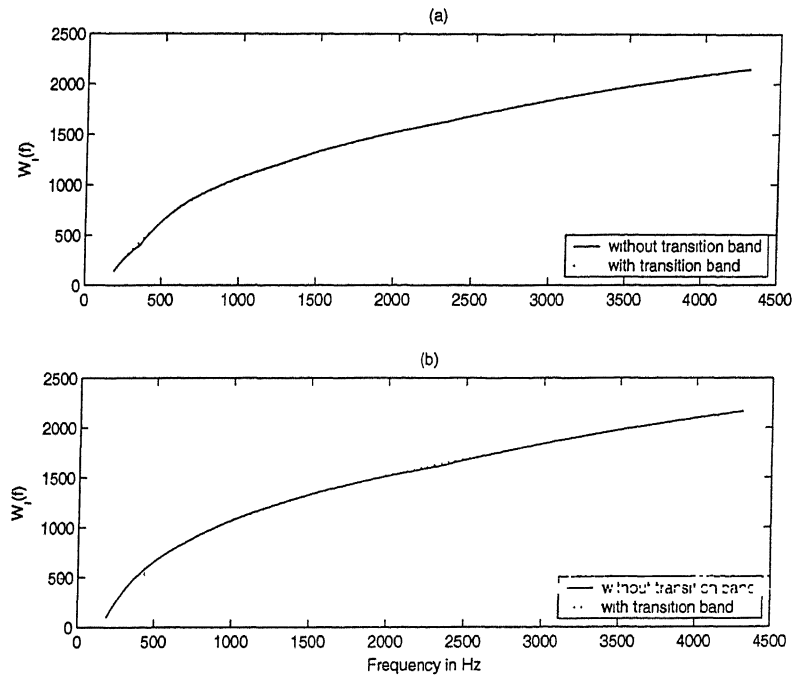
Figure 5.1: Discrete warping function, $W_i(f)$ without transition band and with transition band analysis.

Figure shows the discrete warping function, $W_i(f)$ determined with and without transition band analysis for (a) PnB database and (b) HiL database. Though the curves look very similar, the frequencies at which f-domain is sampled are not exactly the same.

modelling of the non-linearity and the reliability of estimates, resulting in choosing the value of N to be 5. Since the warping function obtained is discrete, we fitted simple curves to it using TableCurve2D to obtain $W(f)$, which is more applicable for continuous spectral patterns. The equations of $W(f)$ for PnB and HiL databases are shown in Table 5.2. Figure 5.2 shows the actual warping function and its closed form approximate, $W(f)$ for PnB and HiL databases. It shows that the curvefits are reliable approximates to their respective originals. Figure 5.3 shows the plot of $W(f)$ for PnB and HiL databases along with mel-warp, log-warp and Stevens & Volkman [18] data points. Mel-warp function defined in Eq. (2.17) is actually a curve-fit to Stevens & Volkman data points, which were the actual mel frequency data points obtained from psychoacoustic studies. The log-warp function refers to simple linear scaling
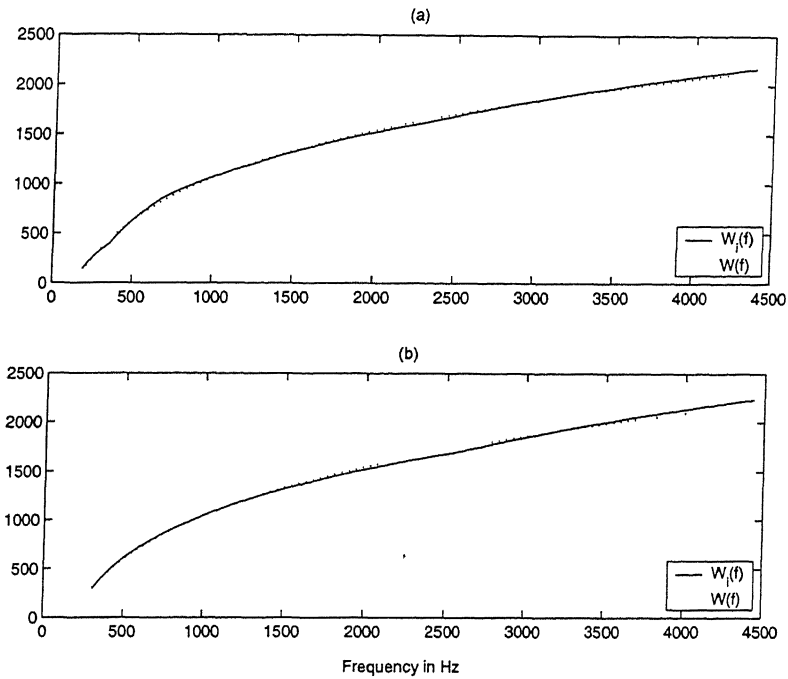
Figure 5.2: Discrete warping function, $W_i(f)$ and its closed form approximate, $W(f)$ *Figure shows the discrete warping function, $W_i(f)$ and its closed form approximate, $W(f)$ given in* Table 5.2 *for (a) PnB database and (b) HiL database. The curvefits to $W_i(f)$ were the best fits obtained from TableCurve2D.*

from the point of view of speaker normalization. The mel scale was derived from psychoacoustic experiments that gave a perceptual measure of pitch. It is a hearing derived scale that relates perceived frequency, and the actual physical frequency. By contrast, the frequency warping function is a speech derived scale that maps physical frequency to an alternate domain, λ, such that in the warped domain the speaker dependencies separate out as translation factors. Note the similarity between $W(f)$ and the mel-warp function at frequencies greater than 3500Hz, and between $W(f)$ and log-warp function at frequencies less than 500Hz. In between these frequencies, $W(f)$ lies between mel-warp and log-warp functions, but closer to log-warp than to mel-warp function. This acts as a compromise between the simple linear scaling (based on speaker normalization) and the mel-scale (based on hearing experiments). This indeed is very interesting that draws some relation between the hearing mechanism and speech production. The degree of vowel normalization provided by $W(f)$ derived
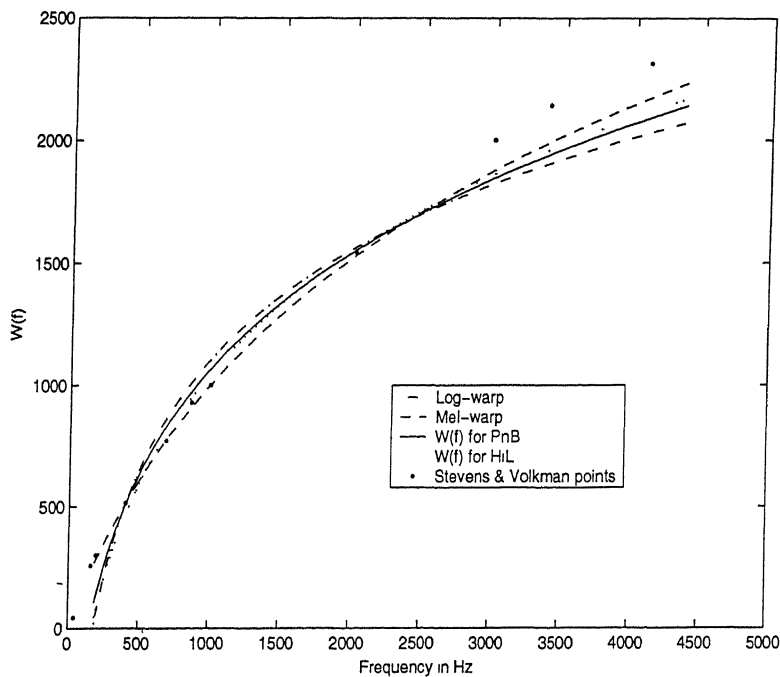
Figure 5.3: Comparison of warping function, $W(f)$, log-warp, mel-warp functions and Stevens & Volkman's actual mel data points.

Figure shows the warping functions for PnB and HiL databases along with log-warp and mel-warp functions. Mel-warp is a function fitted to actual mel data points of Stevens & Volkman. It is interesting to note the similarity of these warping functions, though they are derived from entirely different studies.

from both PnB and HiL databases, compared to log-warp and mel-warp functions is discussed in Section 6.3.

5.5 Summary

The basic theory of scale invariant transformation was presented. A method for incorporating non-linear scaling in such a paradigm was also discussed. The warping function derived out of the study was compared with log-warp and mel-warp functions which was more log-like at frequencies less than 500Hz and more mel-like at frequencies greater than 3500Hz, and acting as a compromise in between these frequencies.

Chapter 6

Comparison of Vowel Normalization Methods in Vowel Classification Performance

In Chapter 4, the efficiency of different vowel normalization methods was discussed in a more qualitative sense. The analytical measures defined in Chapter 4 give an idea of how well the normalization is done by various vowel normalization schemes. But our motivation to study and propose the vowel normalization methods was to make it applicable to speaker independent speech recognition. Hence, the best way to judge the efficiency of these normalization methods would be to implement them on a continuous-density HMM-based recognizer. All the normalization approaches attempt to *normalize* the feature vector of the speech signal, with the intention of reducing inter-speaker differences caused by vocal tract length variations. There are two broad approaches to feature based speaker normalization (1) The first approach is to directly estimate the "gross scale factor, α" either by maximum likelihood (ML) method [3, 4, 11, 16, 25] or by formant estimations (physiological motivations) from the speech data [2, 26] and (2) The second type of systems use a suitable scale-invariant transformation [10], so that there is no need for explicit "α" estimation. Since all the vowel normalization methods that we discussed were based on the formant data of the speakers, it is more logical to judge their normalization performance by applying them on a HMM-based vowel recognizer rather than on a continuous speech recognizer. In this thesis, we have implemented various vowel

normalization schemes on a HMM-based vowel recognizer with two variations, one being to estimate α explicitly in ML sense and the other being using a scale-invariant transformation.

6.1 Hidden Markov Model Based Speech Recognizer

Automatic speech recognizer is a system which allows the computer to recognize the spoken words of a person. Automatic Speech Recognition (ASR) problem is to find a sequence of words to a given set of acoustic features. It involves two stages (1) Feature extraction and (2) Pattern recognition.

Feature-extraction stage is necessary to reduce the dimensionality of the problem and to get a parsimonious representation of the speech signal, where only phonetically relevant information is retained, eliminating unwanted distortions. In our experiments with normalization on the vowel recognizer, we followed two different methods in extracting the features depending on the normalization procedure used. The interested reader is referred to [8, 27] for the details about the feature extraction stage. In brief, standard Davis-Mermelstein [28] filterbank frontend was used to derive Mel Frequency Cepstral Coefficients (MFCC) for the normalization experiments which explicitly estimate the scale factor, α. The normalization experiments with scale-invariant transformation was carried out using the features computed from Weighted Overlapped Segment Averaging (WOSA) [29] analysis with non-uniform DFT.

The pattern recognition problem for automatic speech recognition can be solved in any of the three paradigms (1) Vector quantization (2) Hidden Markov Modelling and (3) Artificial Neural Network. Speech recognition is associated with lot of uncertainties due to different variabilities (like speaker, channel noise, etc.) Stochastic modelling is a flexible method for accounting such variabilities. One of the major advantages of using HMMs in speech recognition problem is their ability to provide a uniform framework for stochastic representation of both acoustic and lexicon rules, along with other sources of knowledge. Eminent works can be referred to for more fundamental details on their usage in ASR [27, 30].

6.2 Speaker Normalization on the Recognizer

As it has been noted earlier that vocal tract size variations which result in scaling of the frequency spectrum of speech signals, account for a major portion of the inter-speaker variations. Hence, it is intuitive to normalize the frequency spectrum of each speaker, with proper estimation of the scaling factor. There exists a class of systems which differ in the manner in which the scale factor is estimated. On the other side, there exists a class of systems which use a suitable scale-invariant transformation thus avoiding the explicit scale factor estimation.

6.2.1 Recognizers With Explicit Scale Factor Estimation

The basic idea in this class of recognizers is to estimate the optimal scaling factor, $\hat{\alpha}$, for every speaker in the training set in maximum likelihood (ML) sense [3, 4], which is used to warp the utterance, thus building a *normalized* HMM. Similarly, during recognition, $\hat{\alpha}$ is estimated for every input speech utterance, which is then used to warp the speech. The decoding of the warped utterance is carried on the normalized HMM. Generally, the scale factor is computed for a speaker with respect to some reference speaker. But in this approach, the reference speaker notion is served by reference HMM. It is clear that the scaling factor estimation process requires a pre-existing HMM model. Hence, an iterative procedure is used to choose the best scaling factor for each speaker and then build a speaker-independent model using the warped training utterance, finally resulting in speaker-normalized model. The interested reader is referred to [3, 4] for full-fledge details about the class of recognizers that estimate the scale factor in ML sense.

6.2.2 Recognizers With Scale Invariance

In this approach, a scale invariant transformation is applied on two scaled (linear/non-linear) spectras, thus transforming them to look similar. The main advantage in this approach is the lack of necessity to estimate the scale factor for every utterance, unlike the method explained in Section 6.2.1. The basic idea is to warp a pair of mutually scaled spectra such that in the warped domain they appear as shifted versions of one another. The magnitude of the Fourier transform of these shifted functions

results in identical feature sets. This method allows the freedom to incorporate the non-linear scaling as a warping function.

Earlier experiments [8] with non-linear scaling function (on a continuous digit recognizer) in this paradigm has shown inferior results compared to linear scaling incorporated in the paradigm explained in Section 6.2.1. Our basic hypothesis is that non-linear scaling should do better than linear scaling as there exists a non-linear relationship between the formant frequencies of speakers. Two obvious reasons can be thought of to explain the nature of the results. One reason may be due to the application of non-linear scaling function derived from vowel data on a continuous digit recognizer, where the consonants may not be normalized. The other reason is that the Fourier transform of the warped spectras are complex quantities. Though the magnitude of the Fourier transforms of two warped spectra removes not only the linear phase, which is basically a speaker dependent term, but also the phase of the Fourier transform of the warped spectra. Thus the reference phase being lost, it is difficult to reconstruct the phase which plays its role in modelling the speech unit. Hence, the scale-invariant transformation approach cannot be applied as it is on the recognizer.

One *smart* but laborious way is to estimate the shift factors in the warped domain in ML sense [31]. Since suitable frequency warping of the uniformly or non-uniformly scaled spectras generates shifted versions in warped domain, the shift factor can be explicitly estimated in ML sense, which finally boils down to the similar class of recognizers with explicit scale factor estimation. But computationally, this is more efficient than the other class as only the shifted versions of the base feature set has to be computed instead of computing the warped spectras for different warping factors.

6.3   Experiments and Results

This section presents an account of the experiments that were carried out to investigate the effectiveness of various speaker normalization procedures in the context of vowels. Speech recognition accuracy was used as a performance measure for speaker normalization.

A Tasks and Database

The data for our vowel recognition experiments was collected from dialect region dr2 of TIMIT database, consisting of 71 male speakers and 31 female speakers. The size of the vocabulary was 12 vowels: /AA/, /AE/, /AH/, /AO/, /AW/, /EH/, /ER/, /EY/, IH/, /IY/, /UH/, /UW/. Inorder to avoid the dialect mismatch, we considered both the training and testing data from dialect dr2 itself. Training and testing sets were separated into male and female data inorder to obtain compact models for both male and female while training the HMM. Training set consisted of 3981 utterances from 53 male speakers and 1770 utterances from 23 female speakers. Testing set was never exposed to models during any part of training. Testing dataset consisted of 1381 utterances from 18 males and 637 utterances from 8 females. These utterances were contributed by all the vowels. After decoding, the number of deletion errors (D), insertion errors (I) and substitution errors (S) were calculated. Percent accuracy, A defined as,

$$A = \frac{N - D - I - S}{N} \times 100\% \tag{6.1}$$

was used to evaluate the performance of various normalization procedures.

B Vowel Recognizer

The experiments were carried out on a HMM based vowel recognition system, using HTK [32]. We conducted our experiments on two different kinds of vowel recognizers (1) Frame-based vowel recognizer and (2) Utterance-based vowel recognizer. In the case of frame-based vowel recognizer, each vowel was modelled by single active state continuous density left-to-right HMM. The observation densities were mixtures of five multivariate Gaussian distributions with diagonal covariance matrices. The basic idea behind this recognizer was to decode each frame separately instead of decoding the whole utterance. Utterance-based vowel recognizer was developed by modelling each vowel by three active state continuous density left-to-right HMM, the observation densities being mixtures of 2 multivariate Gaussian distributions with diagonal covariance matrices.

TIMIT data being recorded over microphone set, is sampled at 16kHz. Speech signals were sectioned with an overlapping window of 20ms frame size and with an overlap of 10ms. A first-order backward difference of of pre-emphasis with factor 0.97 was carried out followed by hamming windowing. 512 point FFT was

A	$\mathcal{M} - \mathcal{M}$	$\mathcal{M} - \mathcal{F}$	$\mathcal{F} - \mathcal{M}$	$\mathcal{F} - \mathcal{F}$
MFCC	45.65	37.08	32.71	47.64
Log-warp	46.91	35.85	34.32	44.96
FDS-PnB	46.91	35.85	34.32	44.96
FDS-HiL	47.51	34.50	35.34	48.54
MBN-PnB	47.17	36.52	34.00	47.98
MBN-HiL	47.17	36.52	34.00	47.98
Mel-warp	46.91	35.85	34.32	44.96
PWB-PnB	47.51	34.50	35.34	48.54
PWB-HiL	47.51	34.50	35.34	48.54

Table 6.1: Recognition performance of various vowel normalization methods on a frame-based vowel recognizer.
Log-warp refers to Eq. (3.16). *FDS-PnB and FDS-HiL are shown in* Table 2.3, *MBN-PnB and MBN-HiL in* Eq. (3.8), *Mel-warp in* Eq. (2.17) *and PWB-PnB, PWB-HiL in* Table 5.2. $\mathcal{M}$ *denotes male gender and* $\mathcal{F}$ *denotes female gender. In the notation* $\mathcal{A} - \mathcal{B}$, $\mathcal{A}$ *denotes the gender of the training data set and* $\mathcal{B}$ *denotes the gender of the test data set.*

taken on each data frame for computing MFCC feature set, while using a 26 channel mel-filterbank. In WOSA based methods, each data frame (without hamming windowing) was sectioned into hamming windowed subframes of 128 samples with an overlap of 90 samples. A smooth spectral estimate was obtained from 255 point autocorrelation function by computing 64-point DFT, the DFT being computed at frequencies determined by the warping function used. Thirteen dimensional feature vectors were used: normalized energy, $c[1] - c[12]$ cepstra which were derived depending on the type of front-end signal processor used in implementing the warping function.

Warping Function Implementation

Consider the function $v = z(f)$, where $z(.)$ is the frequency warping function. Since $z(.)$ warps the scaled spectras to appear as shifted version in v-domain, v-domain is a linear domain. Thus $f = z^{-1}(v)$ gives the discrete frequencies at which the

spectras are to be sampled in f-domain. While computing WOSA based features, DFT is computed at required number of frequencies defined by $f = z^{-1}(v)$. An efficient implementation [33] of non-uniform DFT can be carried out to compute WOSA based features.

Frame-based Vowel Recognizer

In our experiments with frame-based vowel recognizer, we considered only the centre frames of the vowel to model it. This was due to the reason that the vowel will be steady around the centre region rather than at starting and ending instants which are affected by articulations. In a given utterance of a vowel, the first and last frames were excluded and the remaining data was considered in developing the recognizer.

Utterance-based Vowel Recognizer

In our experiments with utterance-based vowel recognizer, the whole utterance of the vowel was considered to model the vowel by three active states, the observation density at each state being mixtures of two multivariate Gaussian densities with diagonal covariance matrices.

C   Vowel Recognition Performance

Inorder to study the effect of warping functions in speaker normalization, we generated gender dependent models by only using the train set data of male speakers or female speakers. The testing was carried out both with and without cross-genders. The experiments on frame-based vowel recognizer were conducted to study the baseline performance of the recognizer. Table 6.1 shows the baseline performances for different warping functions for frame-based vowel recognizer. The normalization experiments were not conducted for frame-based recognizer. Table 6.1 shows that log-warp function, MBN and FDS are consistently better than the other normalization procedures, which confirms our result explained in Chapter 4. The experiments on utterance-based vowel recognizer were conducted to examine the amount of normalization done by various normalization procedures. Table 6.2 shows the baseline recognition performance for different vowel normalization schemes for utterance-based vowel recognizer. Figure 6.1 depicts the percentage improvement in the

(A_b, A_n)	$\mathcal{M} - \mathcal{M}$	$\mathcal{M} - \mathcal{F}$	$\mathcal{F} - \mathcal{M}$	$\mathcal{F} - \mathcal{F}$
MFCC	(60.83, 59.72)	(44.54, 52.53)	(41.76, 47.89)	(59.71, 58.73)
Log-warp	(58.98, 56.84)	(47.96, 53.83)	(40.95, 51.15)	(57.59, 60.36)
FDS-PnB	(58.98, 56.84)	(47.96, 53.83)	(40.95, 51.15)	(57.59, 60.36)
FDS-HiL	(59.28, 56.91)	(46.98, 50.08)	(41.46, 49.59)	(59.54, 59.71)
MBN-PnB	(58.68, 57.28)	(47.47, 53.34)	(41.02, 51.37)	(57.59, 58.73)
MBN-HiL	(58.68, 57.06)	(47.31, 53.18)	(40.87, 51.81)	(56.93, 59.05)
Mel-warp	(58.61, 56.91)	(46.82, 50.57)	(39.54, 46.27)	(59.87, 58.08)
PWB-PnB	(59.28, 56.91)	(46.98, 50.08)	(41.46, 49.59)	(59.54, 59.71)
PWB-HiL	(59.28, 56.91)	(46.98, 50.08)	(41.46, 49.59)	(59.54, 59.71)

Table 6.2: Recognition performance of various vowel normalization methods on an utterance-based vowel recognizer before and after normalization.
Table shows various warping functions which are explained in Table 6.1. *The notation (A_b, A_n) shows that A_b and A_n are the recognition accuracies of baseline (without normalization) and with normalization respectively.*

recognition accuracy with normalization over the baseline for various vowel normalization schemes for cross-gender cases. It is in the case of cross-genders that the normalization performance can be judged more clearly. In similar-gender cases, the normalization was not much effective. It can be clearly seen from Figure 6.1 that MBN does the best normalization for $\mathcal{F} - \mathcal{M}$ case, followed by log-warp function and FDS. This again coincides with the result which we mentioned in Chapter 4 by using some analytical measures. For $\mathcal{M} - \mathcal{F}$ case, MFCC (which actually implements linear scaling) does the best normalization, followed by MBN and log-warp function.

6.4   Summary

A brief overview of the HMM-based vowel recognizer was presented. The implementation of two different vowel recognizers for vowel recognition and normalization task was discussed. The efficiency of vowel normalization methods was studied by applying them to vowel recognition and normalization tasks. MBN, FDS and log-warp
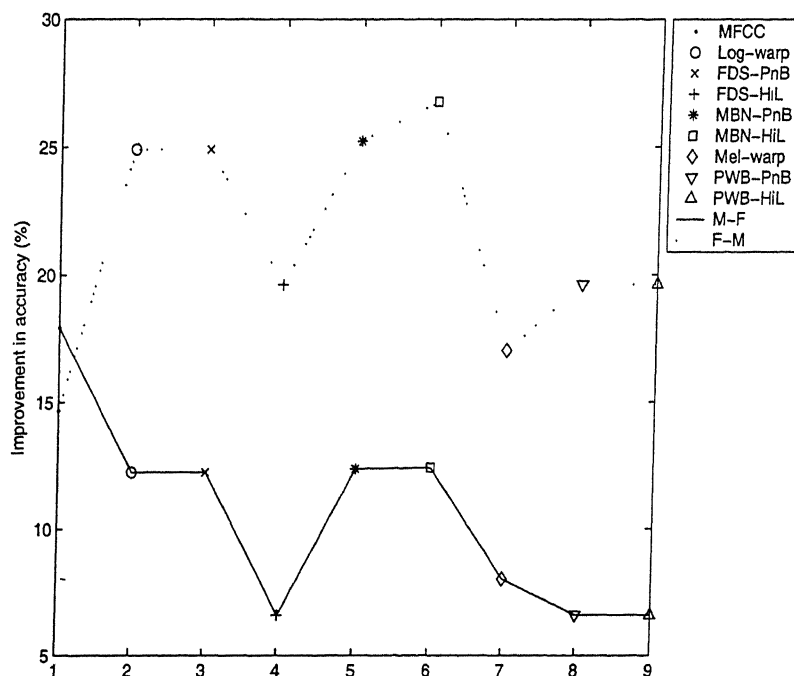
Figure 6.1: Percentage improvement in the recognition accuracy after normalization for various vowel normalization methods on an utterance-based vowel recognizer. *Figure shows the improvement in the recognition accuracy after normalization with respect to the baseline performance for $\mathcal{M} - \mathcal{F}$ and $\mathcal{F} - \mathcal{M}$ cases. The percentage improvement for (A_b, A_n) is calculated as $\frac{A_n - A_b}{A_b} \times 100$, where A_b and A_n are the same as explained in Table 6.2.*

function performed better than other normalization procedures both on frame-based and utterance-based vowel recognizer, thus confirming our previous result obtained through analytical measures.

Chapter 7

Conclusions

In this thesis, we have studied the nature of relationships between formant frequencies of speakers of different age and gender using vowel formant data from Peterson & Barney and Hillenbrand *et al.* databases. Based on this study, a model based non-uniform vowel normalization method is proposed with an aim to achieve robustness to speaker variations in speaker independent speech recognition. We have also made a comprehensive study for frequency dependent scaling method and scale-invariant transformation method and have incorporated them into the state of art recognizers.

Different measures, both objective (viz. residual variance, F-ratio) and subjective (viz. scatter plots) are used in studying the performance of vowel normalization procedures. The best normalization performance in terms of F-ratio and residual variance is obtained for model based normalization procedure. The proposed method gives substantial improvement over simple linear scaling in reducing the variance of vowel clusters. Scatter plots also show similar kind of performance for model based method when compared to other normalization procedures.

The proposed model based normalization method was incorporated into a HMM-based vowel recognizer. From the recognition performance results, it can be inferred that the proposed method does the best normalization for cross-gender cases when compared to other normalization methods.

The frequency-warping necessary to do non-uniform vowel normalization using the proposed model based method turns out to be similar to log-warp whereas for frequency dependent scaling and scale-invariant transformation methods, it turns out to be a compromise between log-warp and mel-warp, closer to log-warp at fre-

quencies less than 500Hz and closer to mel-warp at frequencies greater than 3500Hz. One reason for this kind of complementary behaviour of the warping function, though derived using varied methods but using the same vowel formant data, may be the gross approximation to linear scaling relationship between speakers when the entire frequency range is considered. Since the analysis in frequency dependent scaling and scale-invariant transformation methods is restricted to frequency bands, instead of the whole frequency range, the non-linearities in the speakers are well modelled than in the case of proposed model based method where the analysis is done over the entire frequency range.

Future Work

As it has been already noted, further studies need to be done to understand the contradicting behaviour of the warping functions derived from different vowel normalization methods. Detailed analysis can be done by studying the speaker relationships over different frequency bands. The warping functions for different methods, being derived from vowel formant data, were studied only on HMM-based vowel recognizers. It may be worth trying to implement these methods on a continuous speech recognizer, which gives an idea about the way non-vowels are effected by normalization. If the normalization performance on a speech recognizer turns out to be poor, there opens a wide area of research to normalize the non-vowel sounds.

References

[1] P. E. Nordström and B. Lindblom. A Normalization Procedure for Vowel Formant Data. In *Int. Cong. Phonetic Sci.*, Leeds, England, August 1975.

[2] H. Wakita. Normalization of Vowels by Vocal-Tract Length and its Application to Vowel Identification. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP–25(2):183–192, April 1977.

[3] L. Lee and R. C. Rose. Speaker Normalization Using Efficient Frequency Warping Procedures. In *Proc. IEEE ICASSP'96*, pages 353–356, Atlanta, USA, May 1996.

[4] L. Lee and R. C. Rose. A Frequency Warping Approach to Speaker Normalization. *IEEE Trans. Speech and Audio Processing*, 6(1):49–59, January 1998.

[5] G. Fant. A Non-Uniform Vowel Normalization. Technical report, Speech Transmiss. Lab. Rep., Royal Inst. Tech., Stockholm, Sweden, 1975.

[6] S. Umesh, S. V. Bharath Kumar, M. K. Vinay, Rajesh Sharma, and Rohit Sinha. A Simple Approach to Non-Uniform Vowel Normalization. In *Proc. IEEE International Conference in Acoustics, Speech, and Signal Processing*, Orlando, USA, May 2002. To Appear.

[7] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson. Frequency-Warping in Speech. In *Proc. International Conference on Spoken Language Processing*, Philadelphia, USA, 1996.

[8] M. K. Vinay. Non-Linear Frequency Warping in Speaker Normalization. Master's thesis, IIT, Kanpur, February 2001.

[9] G. E. Peterson and H. L. Barney. Control Methods Used in a Study of the Vowels. *J. Acoust. Soc. America*, 24(2):175–184, March 1952.

[10] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson. Scale Transform in Speech Analysis. *IEEE Trans. Speech and Audio Processing*, 7(1):40–45, January 1999.

[11] T. Kamm, G. Andreou, and J. Cohen. Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability. In *Proc. of the 15th Annual Speech Research S .   .* , pages 175–178, Johns Hopkins University, Baltimore, June 1995.

[12] J. G. Proakis J. R. Deller and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan, New York, 1993.

[13] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.

[14] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler. Acoustic Characteristics of American English Vowels. *J. Acoust. Soc. America*, 97:3099–3111, May 1995.

[15] E. Eide and H. Gish. A Parametric Approach to Vocal Tract Length Normalization. In *Proc. IEEE ICASSP'96*, pages 346–348, Atlanta, USA, May 1996.

[16] P. Zhan and M. Westphal. Speaker Normalization Based on Frequency Warping. In *Proc. IEEE ICASSP'97*, pages 1039–1042, Munich, Germany, April 1997.

[17] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Addison Wesley, New York, 1987.

[18] S. S. Stevens and J. Volkman. The Relation of Pitch to Frequency. *American Journal of Psychology*, 53:329, 1940.

[19] J. B. Allen. How Do Humans Process and Recognize Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, October 1994.

[20] J. Zwicker and E. Terhardt. Analytical Expressions for Critical-Band Rate and Critical Bandwidth as a Function of Frequency. *J. Acoust. Soc. America*, 68(5):1523–1525, November 1980.

[21] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, San Diego, 1990.

[22] L. Cohen. The Scale Representation. *IEEE Trans. Signal Processing*, 41(12):3275–3292, December 1993.

[23] R. A. Altes. The Fourier–Mellin Transform and Mammalian Hearing. *J. Acoust. Soc. America*, 63(1):174–183, January 1978.

[24] S. Umesh, L. Cohen, and D. Nelson. Frequency Warping and The Mel Scale. *IEEE Signal Processing Letters*, 2002. To Appear.

[25] P. Zhan and A. Waibel. Vocal Tract Length Normalization for Large Vocabulary Continuous Speech Recognition. Technical report, School of Computer Science, CMU, Pittsburgh, USA, May 1997.

[26] Y. Ono, H. Wakita, and Y. Zhao. Speaker Normalization Using Constrained Spectra Shifts in Auditory Filter Domain. In *Proc. Eurospeech'93*, pages 355–358, Berlin, Germany, September 1993.

[27] L. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[28] S. B. Davis and P. Mermelstein. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Acoustic, Speech and Signal Processing*, ASSP–28(4):357–366, August 1980.

[29] A. H. Nuttall and G. C. Carter. Spectral Estimation Using Combined Time and Lag Weighting. *Proceedings of the IEEE*, 70(9):1115–1125, September 1982.

[30] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–285, February 1989.

[31] Rohit Sinha and S. Umesh. Non-Uniform Scaling Based Speaker Normalization. In *Proc. IEEE International Conference in Acoustics, Speech, and Signal Processing*, Orlando, USA, May 2002. To Appear.